UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Sistema de información de Gliomas y análisis de expresión diferencial de líneas celulares de Glioblastoma Multiforme

Máster Universitario en Bioinformática y Biología Computacional

Autor: MORENO BOIZA, Vanesa

Tutor: AYUSO SACIDO, Ángel/CARRILLO DE SANTA PAU, Enrique
Departamento de: Fundación de Investigación HM Hospitales/Centro Nacional de Investigaciones Oncológicas

FECHA: Enero, 2018

A mis padres,

gracias a quienes he podido ser y seré.

ÍNDICE

1. ABSTRACT

The Glioblastoma Multiforme (GBM) is the most aggressive brain tumor in the adult, with an average survival of 15 months. It is currently a research object by the HM hospitals research foundation (FiHM), in order to discover biomarkers aiming at becoming therapeutical targets for patient treatment and which in turn can cut down experimental test cost. The Foundation counts on a large volume of data regarding tumor line experiments, otherwise untreatable on a manual basis. Bioinformatics tools have been employed in order to design an information system and a data model with facilitates usage, storage and data retrieval. On the other hand, a differential expression analysis has been carried out among U373, U87 and LN229 glioblastoma tumor lines, by means of Nextpresso pipeline. Different expression outcome discloses mismatches among GBM tumor lines, amounting 4357 differentially expressed genes (DE) between lines U373 and U87, 6007 genes between lines LN229 and U373 as well as 4860 between lines LN229 and U87. Distinct active routes related to angiogenesis have been located too, chiefly in those lines with a lesser proliferation of blood vessels, concurring with the output of previous FiHM experimental trials such as ELISA and inmunofluorescence. Accordingly, we suggest the implementation of further surveys regarding relevant genes for their likely utilization as biomarkers.

2. RESUMEN

El Glioblastoma Multiforme (GBM) se trata del tumor cerebral más agresivo en el adulto, con una supervivencia media de 15 meses. Es objeto de estudio por la fundación de investigación HM hospitales (FiHM) para encontrar biomarcadores que sirvan de dianas terapéuticas para el tratamiento de pacientes y puedan reducir los costes de ensayos experimentales. La fundación dispone de gran cantidad de datos de experimentos de líneas tumorales difícilmente tratables de manera manual. Se han empleado herramientas bioinformáticas para diseñar un sistema de información y un modelo de datos que facilita el acceso, almacenamiento y recuperación de los datos. Además, se ha realizado un análisis de expresión diferencial entre las líneas tumorales de Glioblastoma U373, U87 y LN229, mediante el pipeline de Nextpresso. Los resultados de expresión diferencial revelan diferencias entre las líneas tumorales de GBM con un total de 4357 genes diferencialmente expresados (DE) entre las líneas U373 y U87, 6007 genes entre las líneas LN229 y U373 y 4860 entre las líneas LN229 y U87. También se han encontrado diferentes rutas activas relacionadas con la angiogénesis en mayor medida en las líneas con menor proliferación de vasos sanguíneos, coincidiendo con resultados de previos experimentos de la FiHM como ELISA e inmunofluorescencia. Por lo que sugerimos nuevos estudios sobre genes relevantes de este estudio para un posible uso como biomarcadores.

3. OBJETIVOS

Para la realización del presente proyecto se plantean los siguientes objetivos:

- crear un modelo de datos de un sistema de información de gliomas para la Fundación de Investigación HM Hospitales (FiHM)

- realizar un análisis de expresión diferencial de líneas celulares tumorales de glioblastoma (GBM) con el fin de detectar biomarcadores.

4. INTRODUCCIÓN

El Glioblastoma Multiforme (GBM) es la neoplasia primaria maligna más frecuente del sistema nervioso central (SNC) (Jeffrey, 2014). Se trata del tumor cerebral más agresivo en el adulto, de grado IV, según la clasificación de la Organización Mundial de la Salud (OMS) (Louis, D. N. et al. 2007) que tiene un pronóstico de vida medio de 15 meses tras su diagnóstico (McNamara, M.G. et al. 2013). La presencia de GBM se correlaciona con la presencia de biomarcadores dentro de fluidos biológicos que se pueden obtener a partir de sangre y suero (Skog, J. et al. 2008).
Existe evidencia de una mayor supervivencia libre de progresión del paciente, cuando la terapia se dirige a las vías de señalización del factor de crecimiento en gliomas, lo que demuestra su potencial como objetivo farmacológico para el tratamiento de GBM (Bell, Erica et al. 2011).

El tratamiento convencional para el GBM combina cirugía en los casos en los que la lesión sea accesible sin riesgo de producir daño neurológico grave, radioterapia y quimioterapia. Las terapias actuales no consiguen aumentar la baja supervivencia del paciente.

Los tumores son masas celulares heterogéneas con crecimiento anómalo que causan la invasión y destrucción de órganos y tejidos pudiendo llegar a ocasionar la muerte de un individuo. Una pequeña población de células dentro del GBM denominada Células Madre de Cáncer (CSCs) es responsable de la iniciación y mantenimiento del tumor. El estudio de este tipo de células es de alto interés para facilitar poder abordar su tratamiento y curación.

El laboratorio de investigación para tumores cerebrales de la Fundación de investigación HM hospitales (FiHM) han aislado y caracterizado varios cultivos primarios de células madre de Glioma (GSCs) y realizado diferentes estudios de líneas tumorales de glioblastoma con el fin de detectar biomarcadores objetos de nuevas dianas terapéuticas para el tratamiento de los pacientes.

Realizar este tipo de estudios de investigación de células tumorales es más fácil hoy en día, gracias a las técnicas ómicas y de secuenciación de segunda generación (NGS), que permiten gestionar y analizar grandes cantidades de datos, facilitando obtener nueva información.

El proceso de lectura del DNA se denomina transcripción, de este proceso se obtienen los transcritos (conjuntos de RNA mensajeros) que una vez fuera del núcleo de la célula, en el ribosoma, se traducen para proteínas. Los transcriptomas son muy variables, ya que muestran qué genes se están expresando en un momento dado. Como ya se ha comentado, son particularmente interesantes para los científicos los transcriptomas de las células cancerosas y de las células madre, ya que pueden ayudar a entender los complicados procesos de carcinogénesis y de desarrollo y diferenciación celular.

Inicialmente, en la denominada era pregenómica, el reto era extraer la máxima información posible de muy pocos datos. Se estudiaban unos pocos genes, se intentaban

clonar y estudiar su secuencia para determinar la proteína y ver en qué tejido se expresaba. Este paradigma cambia con la obtención del genoma humano y de otros organismos, lo que supone una revolución en el mundo de la ciencia y la investigación y abre nuevas vías estudio.
Nos encontramos entonces ante un escenario, denominado postgenómico, donde se tienen muchos datos y poca información. Se estudian miles de genes a la vez, sus interacciones con proteínas, el tejido desde punto de vista global, se determinan polimorfismos a lo largo de todo el genoma, etc...

La secuenciación de genomas completos ha permitido conocer las bases moleculares de muchas enfermedades, tanto hereditarias como generadas por mutaciones de novo. Actualmente, los investigadores cuentan con la secuencia completa del genoma humano en las bases de datos electrónicas lo cual permite trabajar con mucha mayor facilidad. Gracias al desarrollo en los últimos años de las tecnologías de secuenciación masiva, partiendo del Proyecto Genoma Humano (1990), ha ido avanzando y expandiéndose el campo de la biología y la bioinformática. El estudio y evolución de las ciencias "Ómicas" es el mejor ejemplo de ello.

El Proyecto Genoma Humano (Human Genome Project, HGP) fue el programa internacional cooperativo de investigación constituido para completar el mapeo y la comprensión de todos los genes de los seres humanos. El conjunto de todos nuestros genes se conoce como nuestro "genoma". Durante el Proyecto Genoma Humano, los investigadores descifraron el genoma humano de tres maneras principales: la determinación del orden, o "secuencia" de todas las bases en el ADN de nuestro genoma; el trazado de mapas que muestran la ubicación de los genes para las principales secciones de todos nuestros cromosomas; y la producción de lo que se denomina "mapas de ligamiento" a través de los cuales los rasgos hereditarios (como los de las enfermedades genéticas) se pueden seguir por varias generaciones.

El Proyecto Genoma Humano reveló que existen probablemente 25.000 genes humanos. La secuencia humana completa ahora puede identificar sus ubicaciones. El resultado del Proyecto Genoma Humano ha brindado al mundo un recurso de información detallada acerca de la estructura, la organización y la función del conjunto completo de genes humanos. Esta información se puede considerar como el conjunto básico de "instrucciones" hereditarias para el desarrollo y funcionamiento del ser humano.
El Consorcio Internacional del Genoma Humano (International Human Genome Sequencing Consortium) se marcó la meta de conseguir secuenciar el genoma humano completo en el periodo 1998-2003 (Collins, F. S. et al. 1998). En el año 2000 anunció casi la completitud de la secuencia (Collins, F. S. et al. 2001), siendo en abril de 2003 cuando completó y publicó la secuencia total.

La secuenciación del RNA con técnicas de segunda generación como RNASeq, hace posible su estudio y permite conocer información muy relevante como lo es el perfil de expresión génica de una célula.
Análisis RNASeq y estudio de expresión diferencial de líneas tumorales puede proporcionar respuestas a las siguientes preguntas.

- ¿Qué relaciones tienen los genes?
- ¿Qué genes cambian significativamente?
- ¿Qué rutas están enriquecidas?
- ¿El perfil génico mejora la predicción de los marcadores clínicos?
- ¿Qué genes se asocian con supervivencia?
- ¿Mejora el diagnóstico?

Es entonces, cuando aparece la necesidad de disponer de procesos automáticos y nuevas herramientas que permitan el manejo de tales cantidades de datos, así como su almacenamiento en repositorios o bases de datos accesibles por la comunidad científica. Aparece una nueva rama, la bioinformática que consiste en la aplicación de tecnologías computacionales a la gestión, análisis de datos biológicos y generación de sistemas de información como repositorios de datos que faciliten una organización y fácil recuperación de la información. Este propósito es de gran importancia para poder unir diferentes resultados de estudios sobre los mismos datos y poder obtener una visión más amplia y conclusiones más certeras.

Los hospitales disponen de una gran cantidad de datos e información de pacientes en diferentes tipos de formatos difícilmente manejables y accesibles por los equipos médicos e investigadores. La información está normalmente desestructurada, repetida y localizada en formatos no fácilmente accesibles lo que conlleva una gran pérdida de tiempo reunir información necesaria para un sólo estudio o proceso.

El éxito en la medicina de precisión depende del acceso a datos genéticos y moleculares de alta calidad de cohortes de pacientes grandes y bien anotadas que acoplan muestras biológicas a datos clínicos integrales, lo que en conjunto puede conducir a terapias efectivas. De tal escenario emerge la necesidad de un nuevo perfil profesional, un bioinformático experto con capacitación en áreas clínicas que pueda dar sentido a los datos multiómicos para mejorar las intervenciones terapéuticas en pacientes, y el diseño de ensayos optimizados de cestas. (Gómez-López, G. et al. 2017).

Como ya se ha comentado, el papel de los sistemas de gestión y repositorios de datos cobra vital importancia en el papel de la ciencia. El conocimiento y la información que pueden extraerse de los datos, son relevante para el mantenimiento de dicho sistema.

Por este motivo, se ha realizado un sistema de información en el que la FiHM pueda albergar la información necesaria y los datos disponibles de sus investigaciones y experimentos en curso o previstos de modo que le sea de gran utilidad a la hora del acceso, almacenamiento y recuperación de la información y favorecer así la obtención de conclusiones resultados del conjunto de sus procesos.

En este caso, los datos que maneja la FiHM son líneas tumorales de glioblastoma, por lo que, además, se ha realizado un análisis de expresión diferencial entre las líneas tumorales de glioblastoma disponibles por la fundación con el objetivo de identificar nuevos biomarcadores.

5. MATERIALES Y MÉTODOS

5.1 HERRAMIENTAS Y MÉTODOS DEL SISTEMA DE INFORMACIÓN

El sistema de información de gliomas ha sido desarrollado siguiendo un modelo de diseño evolutivo en cuanto a la obtención de requisitos de usuario. El diseño del modelo de datos se apoya en los modelos Entidad\Relación y relacional.

- **Modelo de datos**

  Un modelo de base de datos muestra la estructura lógica de la base de datos, incluidas las relaciones y limitaciones que determinan cómo se almacenan los datos y cómo se accede a ellos.

- **Modelo Entidad\Relación**

  El modelo entidad/relación es un modelo de datos conceptual que muestra las entidades relevantes de un sistema de información, así como la forma en la que se distribuyen y relacionan las diferentes tablas del sistema de información.

- **Modelo Relacional**

  Modelo de datos lógico que representa la transformación del diseño conceptual y su normalización para realizar un diseño físico de la base de datos.

Para realizar una implementación del modelo de datos se elige el sistema gestor de bases de datos relacional (RDBMS) que mejor convenga. Para ello, se evalúan las características de varios RDBMS.

- **SQLite**

  Sistema de administración de bases de datos relacionales integrado como una biblioteca en la propia aplicación. La base de datos completa consta de un único archivo en el disco, por lo que es extremadamente portátil. No permite gestión de usuarios para control de acceso. Ofrece una representación de tipos de datos limitados y sólo permite un movimiento de escritura lo que limita el rendimiento. No es recomendable para aplicaciones multiusuario y aplicaciones que requieren altos volúmenes de escritura.

- **PostgreSQL**

  Sistema de gestión de bases de datos relaciones (RDBMS) potente y avanzado, compatible con el estándar SQL de código abierto. Se recomienda su uso cuando la confiabilidad e integridad de los datos son una necesidad absoluta, cuando se pueda requerir integridad con otros sistemas de bases de datos y cuando se trate de diseños complejos. La velocidad es el punto débil de esta base de datos especialmente cuando se trata de accesos de lectura rápida.

- **MySQL**

  Sistema de gestión de bases de datos relaciones (RDBMS)
  - fácil de usar, utilizando el estándar de lenguaje SQL
  - gratis bajo la licencia GPL de código abierto y costo razonable por licencia comercial

- permite su ejecución en muchos sistemas operativos como Linux, Windows...
- disponible en casi todos los proveedores de hosting y de cloud computing, como Amazon WS
- ofrece un soporte técnico ampliamente disponible debido a su popularidad y uso extendido
- soporta bases de datos de gran tamaño
- permite opciones de alta disponibilidad (clúster) y replicación
- escalable y personalizable, permite a los programadores modificar el software para adaptarlo a sus propios entornos específicos
- sistema seguro, implementando funciones de seguridad que proporcionan un acceso confiable a los datos
- rápida, la velocidad es la cualidad más destacada por quienes desarrollan MySQL, función para la que el software fue diseñado principalmente.
- De uso recomendado para sitios y aplicaciones web

Además del RDBMS, se necesita un entorno de desarrollo integrado (IDE) para el acceso físico a la base de datos. La siguiente herramienta proporciona el entorno requerido.

- **MySQL Workbench**

Interfaz de diseño de bases de datos que integra desarrollo de software, administración de bases de datos, diseño de bases de datos, creación y mantenimiento para el sistema de base de datos MySQL.

Para el desarrollo de la aplicación web que de acceso al sistema de datos al usuario, se instala la herramienta Django que utiliza el lenguaje de programación Python.

- **Django**

Framework rápido y versátil de desarrollo de aplicaciones web. Es gratuito y de código abierto, tiene una comunidad próspera y activa, así como una gran documentación y muchas opciones de soporte.
Seguro, permite protección contra algunas vulnerabilidades de forma predeterminada
Escalable, cada parte de la arquitectura es independiente, por lo puede ser reemplazado o modificado si es necesario
Mantenible, diseñado para permitir un código mantenible y reutilizable. Usa modelo vista controlador (MVC)
Portátil, escrito en lenguaje python. Permite ejecutar sus aplicaciones en muchas distribuciones de Linux, Windows y Mac OS X

- **Python**

Python es un potente lenguaje de programación de scripting principalmente diseñado para sistemas Unix. Orientado a objetos, gratuito, multiplataforma, de código abierto y sencillo de sintaxis clara. Ampliamente utilizado, una apuesta por la simplicidad, versatilidad y rapidez de desarrollo ya que, al ser un lenguaje interpretado, no necesita compilar el código fuente para ejecutarlo. Lenguaje que ofrece librerías

específicas de uso bioinformático. Se plantea como interesante opción para realizar todo tipo de programas a ejecutar en cualquier máquina.

El diseño de un sistema de información comprende la creación de unas interfaces gráficas. Para ello las siguientes herramientas permiten realizar el código HTML, relativo a las interfaces de diseño, siguiendo el estilo CSS marcado, a través del IDE Microsoft Visual Studio.

- **CSS (Cascading Style Sheets)**

Hojas de Estilo en Cascada, define el formato de presentación de la información. CSS se utiliza para dar estilo a documentos HTML y XML, separando el contenido de la presentación. Los estilos definen la forma de mostrar los elementos HTML y XML. CSS permite a los desarrolladores Web controlar el estilo y el formato de múltiples páginas Web al mismo tiempo. Cualquier cambio en el estilo marcado para un elemento en la CSS afectará a todas las páginas vinculadas a esa CSS en las que aparezca ese elemento.

- **HTML HyperText Markup Language**

Lenguaje de marcado para hipertextos. Estándar adoptado por los navegadores actuales a cargo del World Wibe Web Consortium (W3C) para la visualización y elaboración de páginas web. Define una estructura básica y un código HTML para la definición de contenido de una página web.
W3C organización dedicada a la estandarización de casi todas las tecnologías ligadas a la web, sobre todo en lo referente a su escritura e interpretación.

- **Microsoft Visual Studio**

Entorno de desarrollo integrado para sistemas operativos Windows. Soporta múltiples lenguajes de programación, tales como C++, C#, Visual Basic

5.2 HERRAMIENTAS Y MÉTODOS DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL

La FiHM ha proporcionado para el análisis de expresión diferencial, los datos de secuenciación en formato RNASeq de las líneas establecidas, tumorales, de origen humano U373, U87 y LN229. La tabla 1 indica una breve descripción de cada una de ellas. Si se desea más información se puede acceder al enlace de las mismas.

LÍNEA CELULAR	DESCRIPCIÓN
U-87 MG	línea celular de glioblastoma primario humano caracterizada por presentar deleciones en la lipofosfatasa
U-373 MG	línea celular continua derivada de un glioblastoma humano
LN-229	línea celular de glioblastoma humano caracterizada por tener mutado p53

Tabla 1. Descripción de líneas celulares utilizadas

Para configurar la herramienta Nextpresso, ha sido necesario consultar el siguiente repositorio, para obtener información del contenido de las distintas firmas moleculares e identificar aquellas necesarias para nuestro estudio. Firmas con las que comparar los resultados de expresión génica. Se añadieron rutas que pudiesen tener relación con angiogénesis para ver qué resultados se obtenían.

- **MySigDB**

Repositorio de firmas moleculares. Versión actual v6.1, contiene 17.786 conjuntos de genes para su uso con GSEA. Se ha de considerar la versión de referencia ya que los conjuntos de genes pueden sufrir cambios o quedar obsoletos en futuras versiones.

Los conjuntos de genes de la base de datos de firmas moleculares (MySigDB) se dividen en 8 colecciones principales, como se muestra en la tabla 2:

ID	DESCRIPCIÓN
H	conjuntos de genes hallmark: firmas expresadas coherentemente derivadas al agregar muchos conjuntos de genes MSigDB para representar estados o procesos biológicos bien definidos
C1	conjuntos de genes posicionales para cada cromosoma humano y banda citogenética
C2	conjuntos de genes curados de bases de datos de pathways en línea, publicaciones en PubMed y conocimiento de expertos en el dominio (BioCarta, KEGG, Reactome, CGP, etc..)
C3	conjuntos de genes motivo basados en motivos conservadores cis-reguladores de un análisis comparativo de los genomas humano, de ratón, de rata y de perro
C4	conjuntos de genes computacionales extráidos de grandes colecciones de datos de microarrays orientados al cáncer
C5	conjuntos de genes GO anotados por términos de Gene Ontology
C6	conjuntos de genes oncogénicos definidos directamente a partir de datos de expresión génica de microarrays de perturbaciones de genes de cáncer
C7	conjuntos de genes inmunológicos definidos directamente a partir de datos de expresión génica de microarrays de estudios inmunológicos

Tabla 2. Colecciones de firmas moleculares del repositorio MySigDB

Se ve que la ruta relacionada con angiogénesis se incluye en la colección h.all.v6.0.symbols (H)

Otras bases de datos secundarias utilizadas y configuradas para el pipeline de Nextpresso son las siguientes.

- **KEGG:** Enciclopedia Kyoto de Genes y Genomas
- **Reactome:** recurso curado de pathways principales y reacciones biológicas humanas
- **Biocarta**: recurso de pathways y funciones biológicas

Las siguientes herramientas se han utilizado para la creación de ficheros de configuración necesarios para la herramienta Nextpresso. Estos ficheros contienen parámetros necesarios para su ejecución.

Este lenguaje es el de edición de los ficheros de configuración de Nextpresso.

- **XML Lenguaje de Marcado Extensible:**

Meta-lenguaje que permite definir lenguajes de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Hace uso de etiquetas para estructurar la información y facilitar su legibilidad.

Esta herramienta ha permitido validar los ficheros de configuración creados XML de Nextpresso.

- **VALIDADORES XML**

Markup Validation Service y XML Validation Program Verifican la validez de marcado de los documentos. Comprueban que el documento XML esté bien formado y su estructura sea la correcta.

Una vez configurados los ficheros, las siguientes herramientas se han utilizado para lanzar la ejecución de Nextpresso.

- **Intérprete de comandos**

La shell del sistema operativo Linux permite interpretar órdenes en lenguaje consola. Se ha utilizado e intérprete de comandos para la ejecución de Nextpresso por Docker.

- **DOCKER**

Docker implementa una API de alto nivel para proporcionar contenedores livianos que ejecutan procesos de manera aislada. Los contenedores son una abstracción en la capa de aplicaciones que combina código y dependencias. Se ejecutan en una sola máquina, comparten el kernel del sistema operativo de la máquina, es decir, se pueden ejecutar varios contenedores en la misma máquina y compartir el núcleo del sistema operativo con otros contenedores, cada uno de los cuales se ejecuta como procesos aislados. Usan menos computación y RAM, ocupan menos espacio que las máquinas virtuales (las imágenes de los contenedores suelen tener un tamaño de decenas de MB vs decenas de GB) y comienzan su ejecución casi al instante. Las imágenes se construyen a partir de capas del sistema de archivos y comparten archivos comunes. Esto minimiza el uso del disco y las descargas de imágenes son mucho más rápidas.
Los contenedores Docker se basan en estándares abiertos y se ejecutan en todas las principales distribuciones de Linux, Microsoft Windows y en cualquier infraestructura, incluidas las máquinas virtuales (VM), bare-metal y en la nube.
Los contenedores acoplables aíslan aplicaciones entre sí y de la infraestructura subyacente. Docker proporciona el aislamiento predeterminado más fuerte para limitar los problemas de la aplicación a un único contenedor en lugar de a la máquina completa.

La figura 1 representa la infraestructura docker.

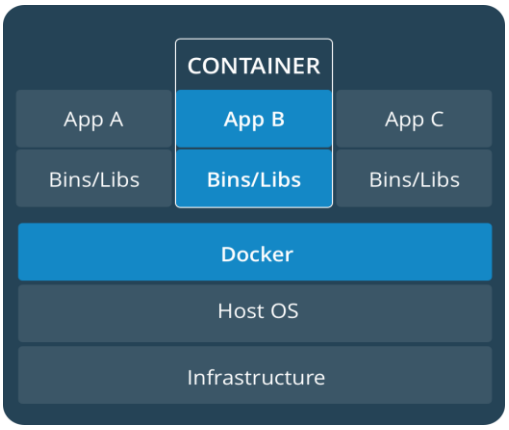


Figura 1. Infraestructura docker

Las herramientas siguientes han permitido revisar el contenido de ficheros R scripts de Nexpresso.

- **R v3.4.0**

Entorno de software libre para gráficos y computación estadística.

- **RSTUDIO v. 0.99.484**

Entorno de desarrollo integrado (IDE) para el lenguaje de programación R. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, depuración y gestión del espacio de trabajo.

La herramienta Nexpresso ha permitido la realización del análisis de expresión diferencial. A continuación, se explica en detalle la herramienta.

Nextpresso v1.9.1 es un pipeline que integra diferentes herramientas, como se puede observar en la figura 2 y realiza, a partir de ficheros de muestras que contienen datos crudos de secuenciación de RNASeq en formato FastQ, un análisis completo de los datos de RNASeq a tráves de cuatro niveles diferentes de ejecución.
1. controles de calidad y contaminación de reads
2. preprocesamiento de lecturas mediante trimming o down-sampling
3. alineación de lecturas frente al genoma o transcriptoma de referencia
4. procesamiento de los alineamientos obtenidos mediante diferentes análisis

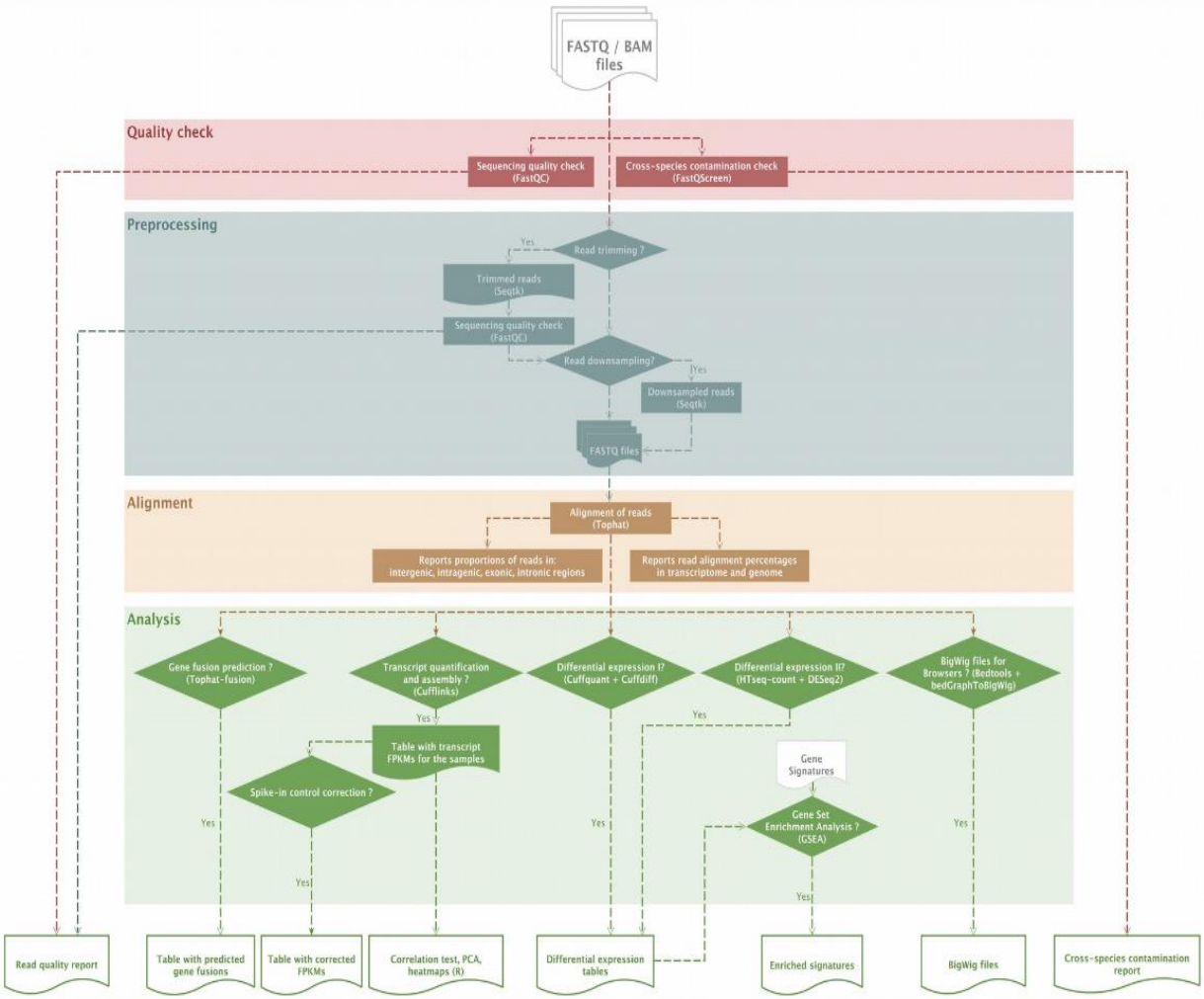


Figura 2. Flujo de trabajo por niveles de ejecución de Nextpresso

A continuación, se describen los pasos de ejecución de Nextpresso para la realización del análisis RNASeq.

1. Calidad de secuencia y control de contaminación (FastQC, FastQScreen)
Antes de proceder al análisis de los ficheros fastq y obtener una conclusión biológica, es importante realizar algún tipo de control de calidad para asegurar que el proceso de secuenciación ha ido correctamente y no hay problemas en los datos que puedan afectar a su posterior estudio. Los secuenciadores normalmente utilizan controles de calidad destinados únicamente a errores producidos por el propio secuenciador.

FastQScreen compara la composición de las muestras FastQ frente a un conjunto de bases de datos de secuencias y produce una salida gráfica con el resumen de mapeo de las secuencias.
FastQC permite llevar a cabo el control de calidad de datos procedentes de secuenciación masiva para detectar fallos en el proceso.

2. Trimming y downsampling (seqtk)

Downsampling (muestreo a la baja) es un método que se usa cuando el número de lecturas de las muestras es muy diferente entre sí. Consiste en ajustar al fichero con menor número de lecturas, de manera aleatoria, de forma que tengamos un número igual o similar. Valores muy diferentes entre las muestras pueden repercutir en el resultado balanceándolo hacia la muestra con mayor número de lecturas.

Seqtk permite hacer downsampling.

En caso de muestras homogéneas y de buena calidad no es necesario realizar el paso 2.

3. Alineamiento (TopHat)

TopHat alinea secuencias cortas de RNA dentro de un genoma de referencia para identificar las uniones de empalme exon-exon, utilizando Bowtie como base. Bowtie es un sistema para alinear cadenas cortas de manera muy rápida con una eficiente gestión de memoria.

TopHat toma como entras el fichero de genoma de referencia, fichero de anotación de referencia gtf y ficheros de lecturas en formato fastq (ficheros con calidad de secuenciación) y devuelve de salida los ficheros de lecturas alineadas en formato BAM y los ficheros de uniones, inserciones y delecciones en formato BED.

4. Ensamblado de transcritos y cuantificación (Cufflinks, Cuffmerge)

Cufflinks ensambla alineamientos de lecturas de RNA en transcritos, cuantifica la expresión calculando estimaciones de su abundancia y realiza expresión diferencial. Normaliza el valor de expresión FPKM (Fragments Per Kilobase of transcript per Million mapped) intra muestra.

Cuffmerge fusiona ensamblados de las muestras con la anotación de un transcriptoma de referencia.

Dado que el siguiente paso es una manera más ajustada de obtener expresión diferencial el paso 4 no se realiza.

5. Expresión diferencial (cuffdiff, cuffnorm)

El análisis de expresión diferencial es un análisis estadístico en el que es muy importante la normalización de los datos para evitar balanceo de resultados. Es necesario tener en cuenta las siguientes asunciones de la normalización:

• Cambios en la expresión son independientes de la abundancia. Las transcripciones raras tienen la misma probabilidad de cambiar en respuesta a un estrés dado que las comunes.

• La mayoría de las transcripciones no se expresan diferencialmente en respuesta a un estrés dado. Relación de expresión típica: tumor/control= 1

• Rango de abundancia comienza en 0. Transcritos negativos = error de medición

• Outliers son biológicamente relevantes (casos en la media son menos interesantes)

En el proceso de expresión diferencial se realiza el cálculo del fold change (ratio de expresión entre dos muestras). El problema es que no tiene en cuenta la variabilidad por lo que es necesario realizar un test estadístico para cada gen y estudiar el correspondiente valor de probabilidad (pvalor). El test estadístico o t-test es el contraste de diferencias en la media entre 2 poblaciones independientes (asume igualdad de

varianzas entre las poblaciones). Otro factor a tener en cuenta es el azar por lo que se ajustan los pvalores obtenidos por múltiple testing mediante control de FDR (tasa de falsos positivos en los resultados).

De acuerdo con el índice de probabilidades y el intervalo de confianza del 95%, el valor de la probabilidad (pvalor) inferior a 0.05, se considera significativo. Por tanto, se consideran diferencias relevantes en aquellos genes cuyo FDR (qvalue) < 0.05.

La figura 3 muestra parte del contenido del fichero resultado de expresión diferencial.

Upregulated in CS101112NGR
Downregulated in CS101112NGR
FDR=0.05

test_id	gene_id	gene	locus	sample_1	sample_2	status	FPKM_1	FPKM_2	log2(fold_change)	test_stat	p_value	q_value	significant
A4GALT	A4GALT	A4GALT	chr22:4	CS123NGR	CS101112NGR	OK	3.5672	30.4483	3.0935	6.0672	0.00005	0.0002724	yes
AADAT	AADAT	AADAT	chr4:17	CS123NGR	CS101112NGR	OK	6.42464	2.24978	-1.51383	-2.65829	0.00005	0.0002724	yes
AASDH	AASDH	AASDH	chr4:57	CS123NGR	CS101112NGR	OK	2.11691	5.28373	1.3196	2.50971	0.00005	0.0002724	yes
ABCA1	ABCA1	ABCA1	chr9:10	CS123NGR	CS101112NGR	OK	13.7428	1.65837	-3.05084	-6.6641	0.00005	0.0002724	yes
ABCA13	ABCA13	ABCA13	chr7:48	CS123NGR	CS101112NGR	OK	0.0991868	4.3197	5.44464	8.23985	0.00005	0.0002724	yes
ABCA2	ABCA2	ABCA2	chr9:13	CS123NGR	CS101112NGR	OK	5.18451	12.5534	1.2758	2.80787	0.00005	0.0002724	yes
ABCA7	ABCA7	ABCA7	chr19:1	CS123NGR	CS101112NGR	OK	1.50679	5.66512	1.91062	3.70241	0.00005	0.0002724	yes
ABCB6	ABCB6	ABCB6	chr2:22	CS123NGR	CS101112NGR	OK	10.1955	27.4742	1.43014	3.25263	0.00005	0.0002724	yes
ABCB9	ABCB9	ABCB9	chr12:1	CS123NGR	CS101112NGR	OK	2.44324	5.5567	1.18543	2.31667	0.00005	0.0002724	yes
ABCC3	ABCC3	ABCC3	chr17:4	CS123NGR	CS101112NGR	OK	3.97057	46.5256	3.55061	6.4205	0.00005	0.0002724	yes
ABCC9	ABCC9	ABCC9	chr12:2	CS123NGR	CS101112NGR	OK	2.43047	0.27643	-3.13625	-4.92363	0.00005	0.0002724	yes
ABCD1	ABCD1	ABCD1	chrX:15	CS123NGR	CS101112NGR	OK	2.22791	13.1125	2.55719	5.11364	0.00005	0.0002724	yes
ABCD4	ABCD4	ABCD4	chr14:7	CS123NGR	CS101112NGR	OK	11.4794	4.86061	-1.23984	-2.66689	0.00005	0.0002724	yes

Figura 3. Resultado de expresión diferencial

Cuffnorm normaliza el valor de expresión a FPKM (Fragments Per Kilobase of transcript per Million mapped) intra e inter muestra. Al tener en cuenta el desbalanceo entre muestras permite un mayor ajuste que Cufflinks. Cuffnorm se procesa por cada comparación realizada y devuelve un fichero con el valor de expresión por cada una de las comparaciones y otro para el conjunto de muestras.

Cuffdiff obtiene los valores de expresión diferencial de cada una de las comparaciones entre muestras. Puede encontrar el detalle de la definición del fichero de formato diff en el ANEXO 1.

En este paso, también se calcula la correlación entre muestras mediante un R script que usa los valores de expresión FPKM y análisis de regresión lineal (coeficiente de correlación de Pearson) para identificar la relación entre dos variables. Considerando valor significativo de aproximadamente 0.9 que implica correlación entre las muestras.

6. Cuantificación y expresión diferencial (Htseq-count, DESeq2)

Htseq-count cuantifica genes y **DESeq2** realiza pruebas de análisis de expresión diferencial. Es una alternativa al uso de Cuffnorm en caso de no haber obtenido por ese método un buen resultado.

7. Visualización en formatos BedGraph y BigWig

BedGraph y BigWig facilitan la visualización gráfica de datos del genoma en servidores como Genome Browser USCS.

BedGraph es un formato track para la visualización de datos continuos. Los ficheros bedgraph no están comprimidos ni son binarios lo que implica mayor esfuerzo a la hora de subir datos al servidor para su visualización.

BigWig es un formato indexado que permite visualizar un gráfico de datos densos y continuos. La indexación permite mostrar una región en particular ya que sólo se transfieren al servidor aquellas secciones de datos necesarias para la visualización. Esto le otorga la ventaja de ser bastante más rápido que otros formatos.

8. Análisis funcional GSEA

Dado que los genes no actúan solos sino en rutas o módulos transcripcionales, el análisis de pathways puede aportar información relevante.

GSEA (Gene Set Enrichment Analysis) es un método computacional que determina si un conjunto de genes definido a priori muestra diferencias estadísticamente significativas entre dos estados biológicos.
Dado que un análisis de expresión diferencial puede no tener genes diferencialmente expresados, GSEA supone que pequeñas variaciones individuales de los genes pueden dar lugar a una diferencia significativa a nivel de ruta, en cambio otros métodos como ORA (Over Representation Analysis) se ven limitados al usar una lista de genes diferencialmente expresados, debido a esto, GSEA puede aumentar la potencia del análisis.

La figura 4 representa un esquema del modelo GSEA.

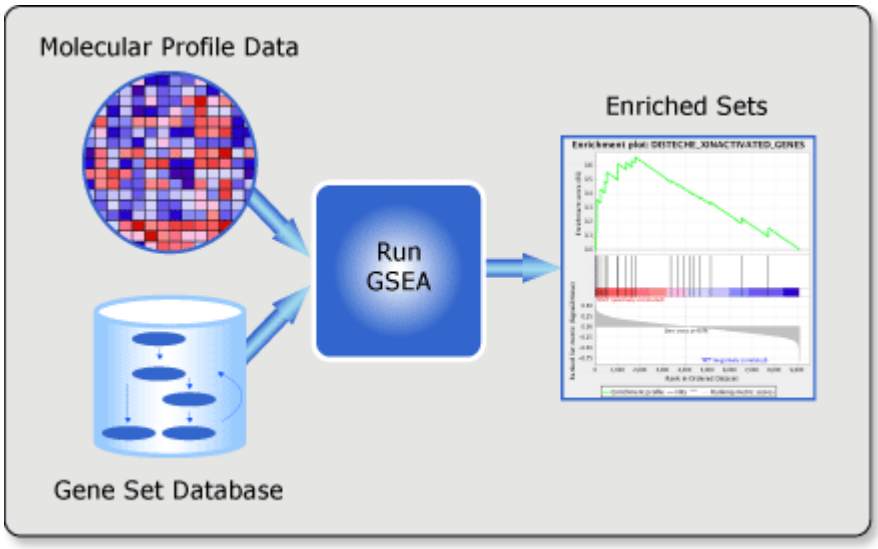


Figura 4. Modelo GSEA

Ciclo de proceso
1. Cálculo de ES (Enrichment Score).
2. Estimación de significancia.
3. Cálculo NES: (Normalized Enrichment Score). No es posible comparar conjuntos de diferente tamaño.
4. Ajusta por Multiple testing.

El informe de análisis resultante considera relevantes de estudio aquellos conjuntos de genes con un FDR < 25% como los más propensos a generar hipótesis interesantes e impulsar investigación (en caso de < 7 réplicas puede utilizarse FDR < 0.05)

La tabla 3 contiene los formatos de datos de entrada a GSEA.

Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, or pcl	You create the file.
Phenotype labels	Contains phenotype labels and associates each sample with a phenotype.	cls	You create the file or have GSEA create it for you.
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You export gene sets from the Molecular Signature Database (MSigDb) or create your own.
Chip annotations	Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.	Chip	You use a chip file installed with GSEA, download a chip file from the GSEA web site, or create your own chip file.

Tabla 3. Formatos de datos de entrada

Nextpresso utiliza GSEA Prerranked, recomendable en casos de análisis funcional de RNASeq que normalmente cuentan con un número inferior a 7 replicados.
El análisis se realiza a partir de los datos de expresión diferencial obtenidos por cada una de las comparaciones realizadas, de forma que, se crea una lista de genes ordenados por log2(fold change) a la que se comparan conjuntos de genes específicos (pathways) tomando como valores de interés los extremos de la lista.

La figura 5 muestra un reporte de resultados de GSEA.

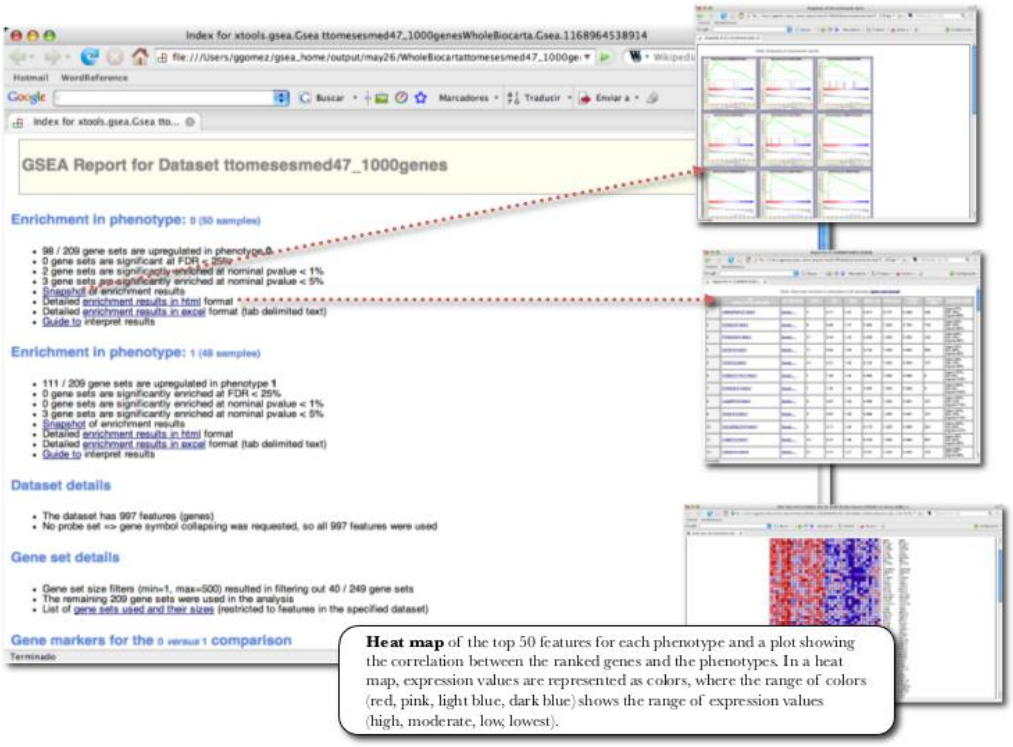


Figura 5. Reporte GSEA

9. Predicción de fusión de genes (TopHat-Fusion)

TopHat-Fusion es un algoritmo diseñado para descubrir transcripciones que representan fusión de genes. El proceso de fusión puede ocurrir debido a una traslocación, delección intersticial o inversión cromosomal. Es una versión mejorada de TopHat, alinea lecturas RNA-seq sin depender de la anotación existente lo que le permite descubrir productos de fusión derivados de genes conocidos, desconocidos y variantes de genes conocidos no anotados.

El estudio de fusión de genes tiene sentido realizarlo cuando se trata de análisis Paired End en los que se secuencian los dos extremos de la molécula.

Para interpretar los resultados del análisis de expresión diferencial, se han utilizado diferentes repositorios de datos biológicos que han facilitado información específica, de forma que, se ha podido comparar la información entre distintas bases de datos y ampliarla con el conjunto de información de todas ellas. A continuación, se detalla el uso de cada uno de ellos.

- **HGNC**

Hugo Gene Nomenclature Comittee del Instituto Europeo de Bioinformática (EMBL-EBI). Organismo responsable de aprobar y proporcionar identificadores unívocos para loci humanos, incluidos genes que codifican proteínas, genes ncRNA y pseudogenes, para permitir una comunicación científica inequívoca. Proporciona un repositorio de información al alcance del usuario que facilita la obtención de información de identificación de genes. Permite consultar información de sinónimos de genes.

Esta herramienta ha permitido obtener sinónimos y nomenclatura de los genes del estudio.

- **GeneCards**

Base de datos que proporciona información completa y fácil de usar sobre todos los genes humanos anotados y predichos. Integra automáticamente los datos centrados en genes de ~ 125 fuentes web, incluida información genómica, transcriptómica, proteómica, genética, clínica y funcional.

Esta herramienta ha permitido obtener sinónimos, nomenclatura de genes y código identificador ENSG de genes en estudio.

- **UNIPROT**

Universal Protein Resource (UniProt). Recurso integral para la secuencia de proteínas y datos de anotación. Es una colaboración entre el Instituto Europeo de Bioinformática (EMBL-EBI), el Instituto Suizo de Bioinformática y el Recurso de Información de Proteínas (PIR). Permite la consulta de código ENSG, alias e información específica de cada gen.

Esta herramienta ha permitido obtener sinónimos, nomenclatura de genes y código identificador ENSG de genes en estudio.

La herramienta PUBMED se ha utilizado para buscar información bibliográfica y literatura asociada que apoye los datos de resultados y definiciones relevantes del estudio en curso.

- **PUBMED**

Base de datos gestionada por el NCBI, de acceso libre y especializada en campos de la biomedicina y la salud, con más de 27 millones de referencias bibliográficas de literatura biomédica de MEDLINE (principal base de datos bibliográficos de la Biblioteca Nacional de Medicina de EE. UU. (NLM)), revistas de ciencias de la salud y libros en línea que cubren partes de las ciencias de la vida, las ciencias del comportamiento, las ciencias químicas y la bioingeniería. Proporciona acceso a sitios web relevantes adicionales y enlaces a otros recursos de biología molecular del NCBI.

Una vez obtenidos los resultados GSEA, se revisaron las firmas moleculares en el repositorio de datos MySigDB para identificar el contenido de las siguientes rutas

Las pathways o rutas, definidas en la tabla 4, se encuentran en la colección h.all.v6.0.symbols (H).

ID	DESCRIPCIÓN
HYPOXIA	Nombre sistemático M5891 Genes regulados en respuesta a bajos niveles de oxígeno (hipoxia)
ANGIOGÉNESIS	Nombre sistemático M5944 Genes regulados por incremento durante la formación de vasos sanguíneos (angiogénesis)
APOPTOSIS	Nombre sistemático M5902 Genes que median la muerte celular programada (apoptosis) por activación de caspasas
P53_PATHWAY	Nombre sistemático M5939 Genes involucrados en las rutas y redes de p53.

Tabla 4. Descripción de algunas rutas relacionadas con angiogénesis.

Para visualizar el resultado de expresión diferencial con una gráfica HeatMap, se necesitan configurar unos ficheros con los datos filtrados. Hacemos uso del intérprete de comandos para generar estos ficheros.

La herramienta Morpheus se ha utilizado para visualizar el HeatMap de los ficheros de expresión diferencial en formato gct.
Morpheus es una herramienta web que permite la visualización mediante gráfico HeatMap de un fichero de datos en formato gct.

6. RESULTADOS

6.1 RESULTADOS DEL SISTEMA DE INFORMACIÓN

El primer paso para la realización del proyecto es configurar el equipo de trabajo con las herramientas necesarias. Se utiliza un equipo con distribución Ubuntu y sistema operativo Linux. Se configura workbench, django, docker y MySQL server.

En base a la finalidad principal del sistema de información del proyecto BrainfGM, permitir al usuario el acceso y almacenamiento de la información de una manera organizada y abstracta de modo que facilite la obtención de información relevante de los datos para el investigador, se realiza un análisis sobre cuáles serían los principales requisitos a cubrir tanto por el modelo de datos como por la aplicación web propuestos para el sistema de información de gliomas.

El modelo de datos y la aplicación web propuestos presentan información exclusiva del sistema de información de gliomas del proyecto BrainfGM y han sido diseñados según las necesidades específicas de la FiHM.

Modelo Entidad\Relación

El modelo Entidad/Relación de la figura 6, representa el diseño del sistema de información de gliomas y cubre principalmente la necesidad de almacenamiento de información relativa a los estudios realizados con las distintas líneas celulares de glioma y resultados de análisis obtenidos de diferentes formatos en los que se representan.

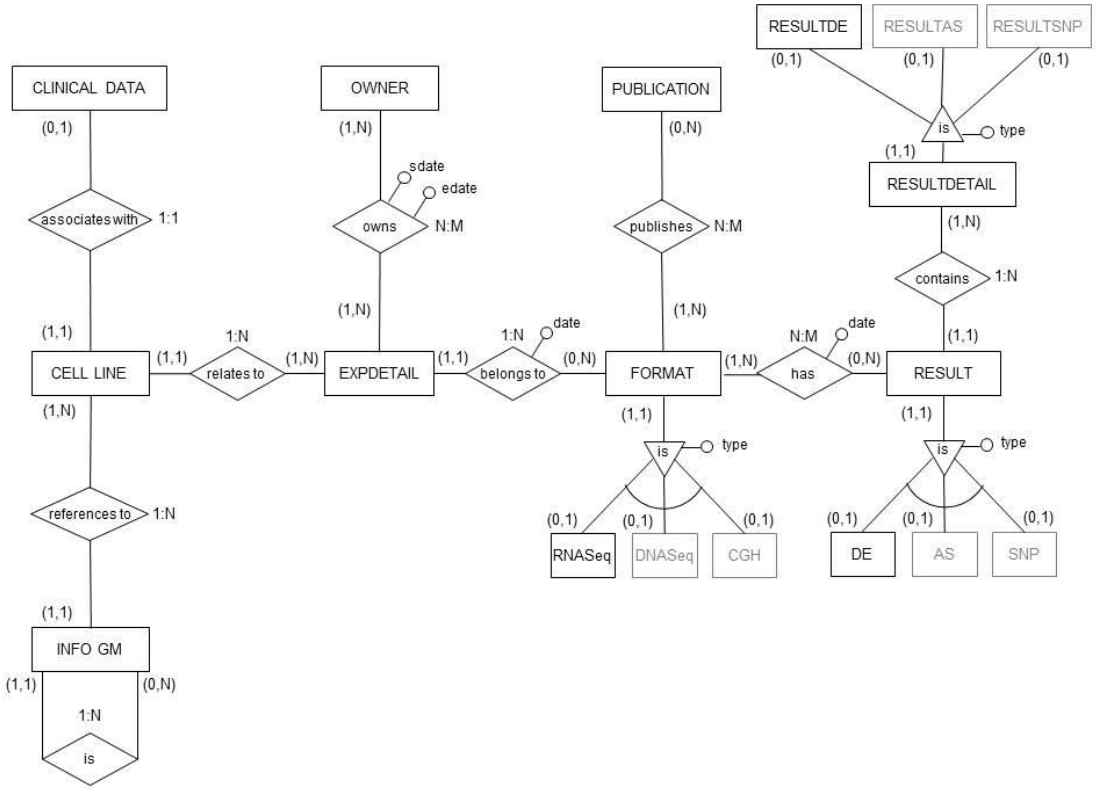


Figura 6. Modelo Entidad/Relación del sistema de información de gliomas

El modelo relacional, así como el detalle de elementos del modelo de datos se encuentran en ANEXO 2.

Comparativa de gestores de bases de datos relacionales.
Se realiza un estudio de las características de RDBMS para seleccionar el que mejor convenga para la realización del proyecto. En la tabla 5 se presenta la información relevante de este estudio y se marca con un check en aquellos RDBMS en los cuales la característica en sí, es un valor resaltado. No contener un check no significa que el RDBMS no disponga de dicha característica. Por las características generales que presenta y el soporte para web, se considera la mejor elección el uso del RDBMS MySQL.

Características\RDBMS	SQLite	MySQL	PostgreSQL
Potencia		✓	✓
Velocidad		✓	
Seguridad		✓	✓
Portabilidad	✓		
Rango de tipos de datos		✓	✓
Rendimiento		✓	✓
Soporte a diseños muy complejos			✓
Soporte web		✓	✓
Soporte técnico y documentación		✓	

Tabla 5. Comparativas de RDBMS

Especificaciones del sistema de información

A continuación, se presenta una relación del análisis a alto nivel de los principales requerimientos del sistema de información para la implementación de una aplicación web que cubra las necesidades del modelo de datos y encapsule procesos para aportar información relevante al investigador, dados los datos almacenados en ella. Se ha de tener en cuenta que el usuario final de la aplicación serán principalmente científicos investigadores.

Preguntas que se plantean para que la aplicación web de respuesta en función del modelo de datos definido:

1. Obtener información de los propietarios por fecha de un detalle experimento
2. Obtener los pases y replicados que forman parte de un experimento
3. Obtener ordenados por fecha los tipos de formatos disponibles para un detalle experimento concreto
4. Obtener información de los datos clínicos asociados a una determinada línea celular
5. Dada una línea celular obtener información de glioma asociada

6. Dada una línea celular obtener los detalles de experimentos asociados ordenados por fecha descendiente
7. Obtener información de las publicaciones relativas a un formato
8. Obtener información de los ficheros de resultados asociados a un formato
9. Obtener los detalles de resultados de un fichero de resultados de expresión diferencial concreto
10. Dado un gen o lista de genes relativos a un detalle de experimento, detectar en qué comparaciones está sobreexpresado
11. Gen/genes sobreexpresado up/down en una determinada comparación
12. Obtener un gráfico tipo boxplot ordenando las líneas celulares de mayor a menor expresión de un gen dado por el investigador

Las clases se definen en función del modelo de datos y las necesidades propias del sistema de información para dar respuesta a las preguntas planteadas.

Las principales clases del sistema se corresponden con entidades del modelo de datos: InfoGM, CellLine, ClinicalData, ExpDetail, Owner, Format, Publication, Result, ResultDE y una clase Utilities añadida para albergar procesos complejos de tratamiento de información necesarios para dar respuesta a las preguntas planteadas.

Clases principales, se corresponden con las entidades del modelo datos y van a contener los siguientes métodos principales.

- método constructor de la clase: permite la instancia de un objeto de la clase
- una propiedad por cada atributo de la entidad que permitan las acciones get y set: permite obtener y almacenar valores por cada atributo de la entidad a la que corresponde.
- métodos de creación, eliminación y modificación de un objeto de la clase.

Clase Utilities, va a contener:

- método constructor de la clase
- métodos necesarios para manipular datos de las distintas clases e implementar funcionalidad compleja como visualización de gráficos, procesamiento de información de distintas entidades, procesos de cálculos matemáticos y/o estadísticos sobre los datos.

La aplicación web deberá permitir el alta, baja y modificación de los datos al usuario con perfil de investigador. Para el alta de resultados de detalles de análisis realizados existe una excepción. La gran cantidad de datos de resultados hace inviable su manipulación manual por lo que es necesario un proceso de carga masiva de datos. Este proceso será un módulo de servidor. Como los resultados vienen almacenados en ficheros y su análisis puede haberse realizado por diferentes tecnologías y empresas hay que tener en cuenta que pueden tener diferente formato y contenido por lo que el proceso de carga deberá apoyarse en una configuración de correspondencia entre la información del fichero y los campos del modelo de datos, por ejemplo, mediante el uso de ficheros de configuración XML.

Además, se propone que la aplicación permita la opción de ofrecer los resultados de los análisis de forma gráfica para facilitar la comprensión por parte del investigador.

Para la implementación de las aplicaciones se utilizará lenguaje Python. El desarrollo de la aplicación web se apoyará en el uso de Django.

La aplicación web deberá de tener en cuenta un control de usuarios de acceso, para ello deberá de apoyarse en un modelo de datos que cuente con las siguientes entidades:

- USER: entidad que almacena información de los usuarios del sistema. Su clave principal será el identificador unívoco de usuario (iduser).
- ROLE: entidad que almacena información de los permisos de acceso al sistema de información. Su clave principal será el identificador unívoco de role (idrole).
- USER-ROLE: entidad proveniente de la interrelación de cardinalidad (N:M) entre las entidades USER e USER-ROLE que almacena la relación de permisos de un usuario. La clave principal estará formada por las claves ajenas iduser e idrole.

Interfaces del sistema

A continuación, se muestran unas interfaces tipo, de la aplicación web, del sistema de información de gliomas.

Interfaz Acceso

La figura 7 representa el menú de acceso a la aplicación donde se solicitan los datos de usuario y password para realizar un control de acceso, de forma que, sólo tengan acceso los usuarios habilitados para tal fin.

Figura 7. Interfaz acceso

Interfaz Opciones

La figura 8 representa el menú principal con las distintas opciones del sistema. Cada una de las opciones llevará a su vez a un menú individual por elemento seleccionado. Por ejemplo, al seleccionar Cell Line accederá a una interfaz del mismo formato donde podrá dar de alta, baja, consultar o modificar una nueva línea celular.

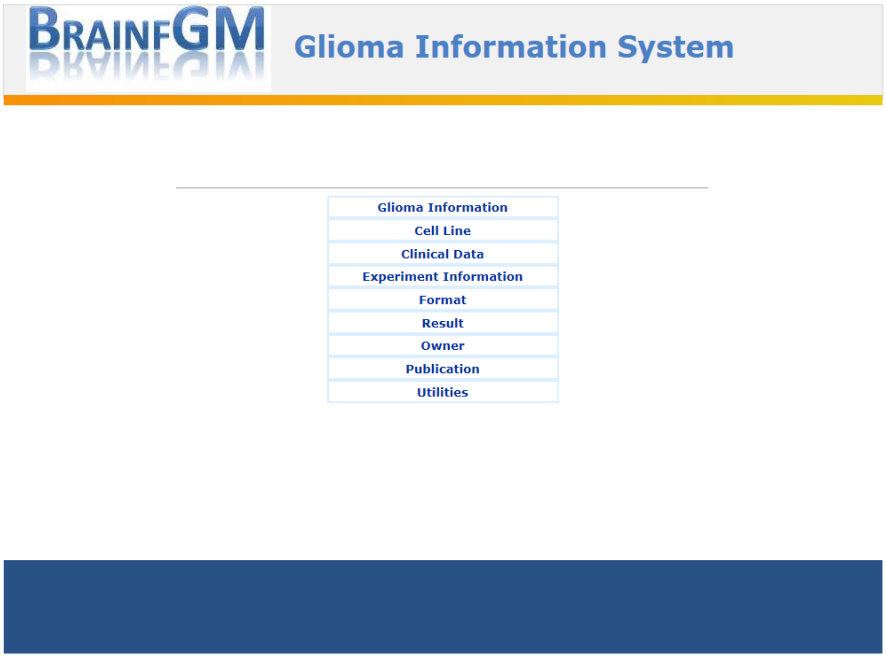


Figura 8. Interfaz opciones

Interfaz Contenido

La figura 9 representa el menú de un tipo de contenido, en este caso, se trata de la opción New Cell Line dentro de la opción Cell Line del menú principal para dar de alta una nueva línea celular en el sistema. Cada contenido mostrará una información específica que dependerá del elemento principal seleccionado.

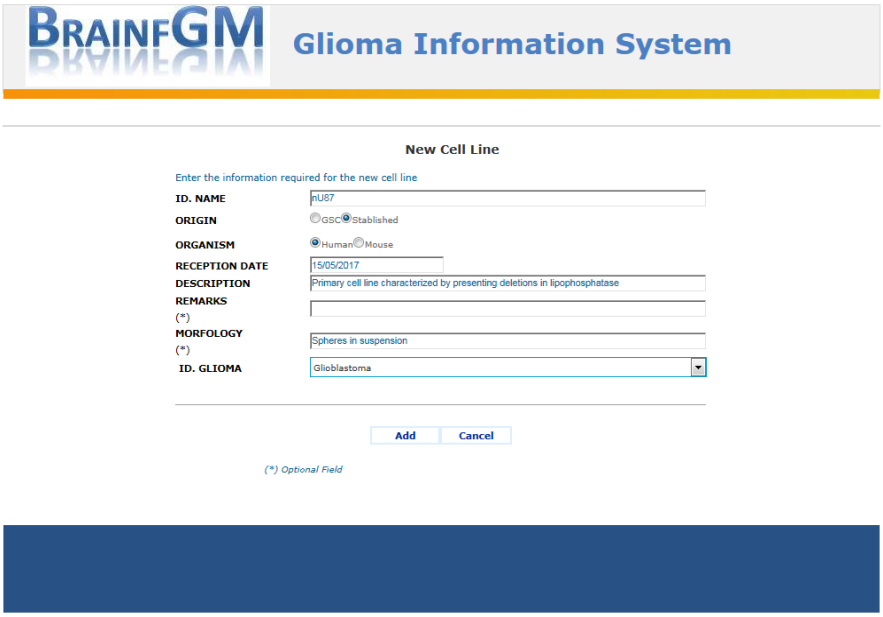


Figura 9. Interfaz contenido

6.2 RESULTADOS DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL

La ejecución del pipeline de Nextpresso se ha lanzado por medio de Docker, lo que ha permitido un mayor encapsulamiento.

Comparativa RNASeq vs otras tecnologías
Se realiza una comparativa entre diferentes tecnologías ómicas y de segunda generación (NGS) para verificar la calidad de la herramienta de procedencia de los datos de secuenciación. A continuación, se muestra la tabla 6 con el resultado de la comparativa entre RNASeq y diferentes tecnologías de secuenciación. Debido a sus características generales, se decide que la técnica RNASeq es la mejor elección, principalmente, basándonos en la alta calidad de su secuenciación.

Tecnología	**Microarray**	**Secuenciación de cDNA o EST**	**RNA-Seq**
Especificaciones tecnológicas			
Principio	Hibridación	Secuenciación de Sanger	Secuencia de alto rendimiento
Resolución	De varios a 100 pb	Base individual	Base individual
Rendimiento	Alto	Bajo	Alto
Confianza en la secuencia genómica	Sí	No	En algunos casos
Ruido de fondo	Alto	Bajo	Bajo
Solicitud			
Mapeo simultáneo de regiones transcritas y expresión génica	Sí	Limitado para la expresión génica	Sí
Rango dinámico para cuantificar el nivel de expresión génica	Hasta unas pocas veces	No practico	> 8,000 veces
Capacidad de distinguir diferentes isoformas	Limitado	Sí	Sí
Capacidad de distinguir la expresión alélica	Limitado	Sí	Sí
Cuestiones prácticas			
Cantidad requerida de ARN	Alto	Alto	Bajo
Costo por mapeo de transcriptomas de grandes genomas	Alto	Alto	Relativamente bajo

Tabla 6. Comparativa de tecnologías ómicas Arrays vs RNASeq

Para lanzar el análisis de RNASeq a través de Nextpresso, es necesario configurar los ficheros de parámetros que el pipeline necesita. Estos ficheros son específicos para el análisis comparativo realizado con las tres líneas celulares U373, U87 y LN229. La tabla 7 muestra la correspondencia del nombre de la línea celular con el nombre asignado a la misma para la comparación:

LÍNEA CELULAR	NOMBRE EN ANÁLISIS	DETALLE
U373	CS101112NGR	Línea celular establecida (CS) 101112: correspondiente a los números de replicados 10,11 y 12 NGR: iniciales del propietario Noemí García Romero
U87	CS123NGR	Línea celular establecida (CS) 123: correspondiente a los números de replicados 1, 2 y 3 NGR: iniciales del propietario Noemí García Romero
LN229	CS192021NGR	Línea celular establecida (CS) 192021: correspondientes a los números de replicados 19,20 y 21 NGR: iniciales del propietario Noemí García Romero

Tabla 7. Correspondencia entre líneas celulares y nombres de análisis

Nextpresso devuelve diferentes carpetas en las que almacena la información de ficheros intermedios y resultantes de cada uno de los procesos. Se revisan las salidas de Nextpresso para todas las muestras. A continuación, se explica el resultado de una de ellas.

Se comienza la supervisión de resultados de Nextpresso revisando la calidad de todas las muestras.

Resultado tipo de control de calidad de secuencia y contaminación realizado con FastQScreen

La gráfica de la figura 10 visualiza que la composición de la muestra mapea principalmente con genoma humano.

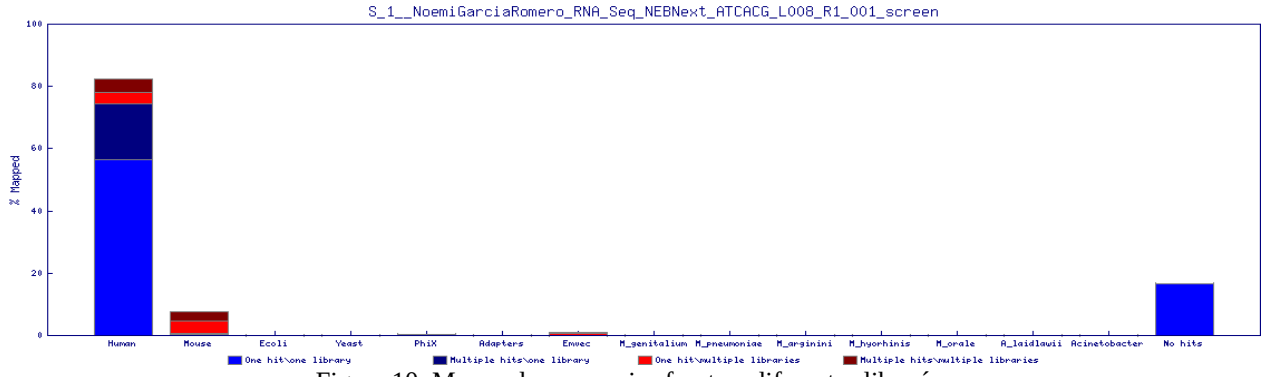


Figura 10. Mapeo de secuencias frente a diferentes librerías

Como conclusión, los resultados de las gráficas de FastQScreen de mapeo de secuencias frente a componentes de distinta naturaleza, exhiben que las muestras mapean principalmente con componente de genoma humano.

Resultado tipo de control de calidad de secuencia y contaminación realizado con FastQC

Se revisan todos los pasos de control de calidad y secuencia. Se considera que las muestras son de alta calidad. A continuación, se explican los pasos más relevantes y el resultado de ellos.

La figura 11 nos da información general del resultado FastQC de la muestra y la figura 12 información general del archivo: nombre, tipo, programa de secuenciación, número total de secuencias, secuencias marcadas con mala calidad, tamaño de secuencia y porcentaje de C y G

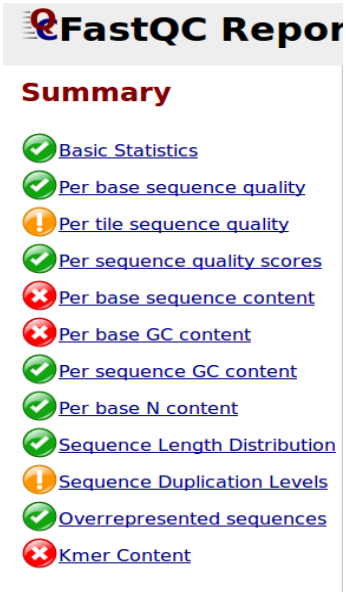
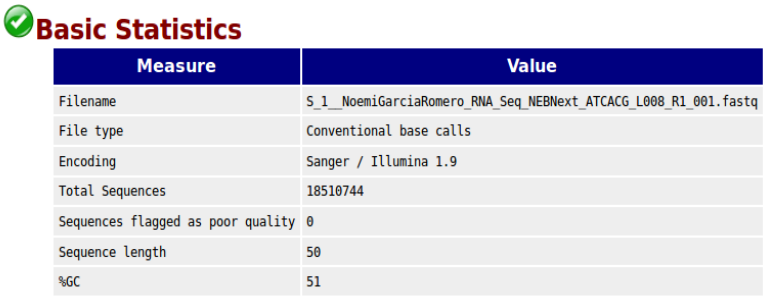


Figura 11. Información general del resultado FastQC



Measure	Value
Filename	S_1__NoemiGarciaRomero_RNA_Seq_NEBNext_ATCACG_L008_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	18510744
Sequences flagged as poor quality	0
Sequence length	50
%GC	51

Figura 12. Información general archivo

La figura 13 refleja la calidad de la secuenciación mediante el valor de Q para cada una de las bases de todas las secuencias. El valor de Q para cada base puede ir de 0 a 40, considerándose aceptable la secuenciación para valores de Q iguales o mayores a 25. El amarillo refleja los valores de Q para los cuartiles del 25 y 75 %, la línea azul representa la media de los valores de Q asignados a cada base y la roja la mediana.

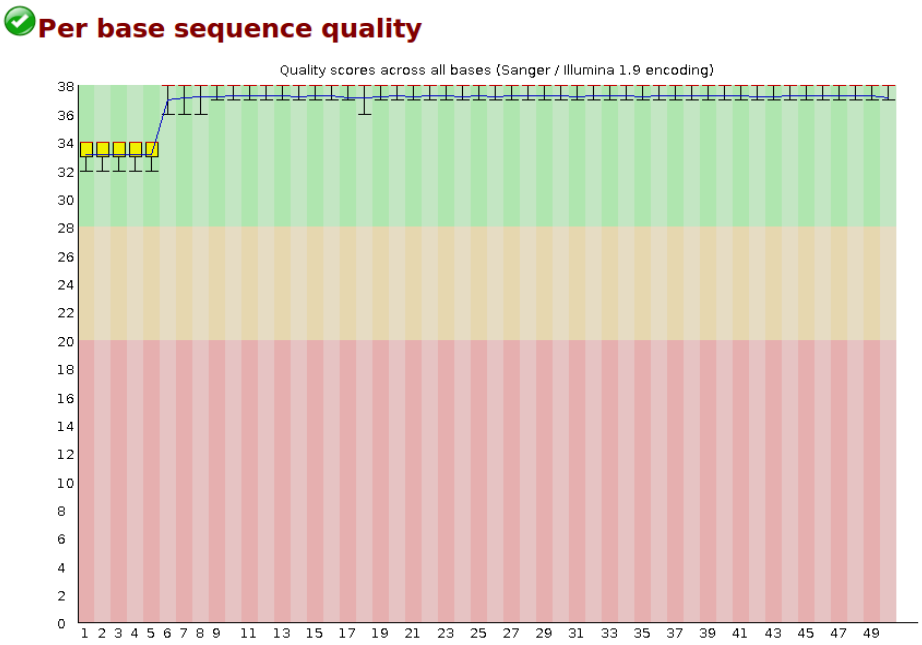


Figura 13. Calidad de secuencia por base

Como conclusión, los valores FastQC de calidad de secuenciación por base de las muestras en un rango de 0 a 40, resultan cercanos al máximo lo que denota buena calidad.

La figura 14 nos informa de si la reacción se ha producido correctamente dentro del canal o célula de flujo (flow cell). Dentro del canal la reacción se da cuando se pasan diferentes reactivos por el surco que lo atraviesa. Si la lectura se da sin interferencias, se muestra el fondo azul por completo. Las manchas indican interferencias en el proceso que puedan ser debidas a burbujas en la reacción por no desgasificar los reactivos, a suciedad, etc...

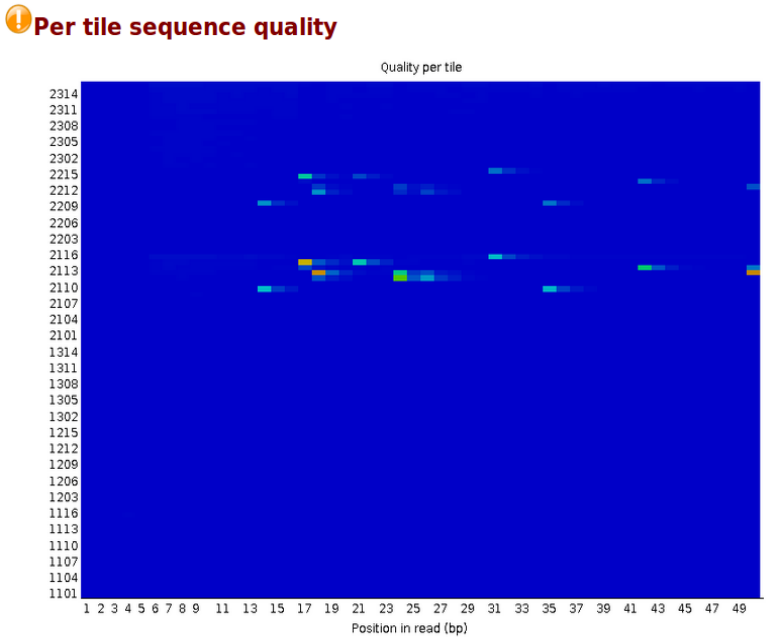


Figura 14. Calidad de secuencia por canal

Como conclusión, según las figuras FastQC de calidad de secuencia por canal, aparecen en algunas de las muestras interferencias no muy significativas en el proceso de secuenciación.

La figura 15 indica la proporción de calidad de las secuencias. La mayoría de secuencias tienen valores altos, no se aprecian conjuntos de datos de baja calidad. Esta representación puede ser de utilidad para predecir el número de secuencias a descartar si se limita el valor de calidad Q.

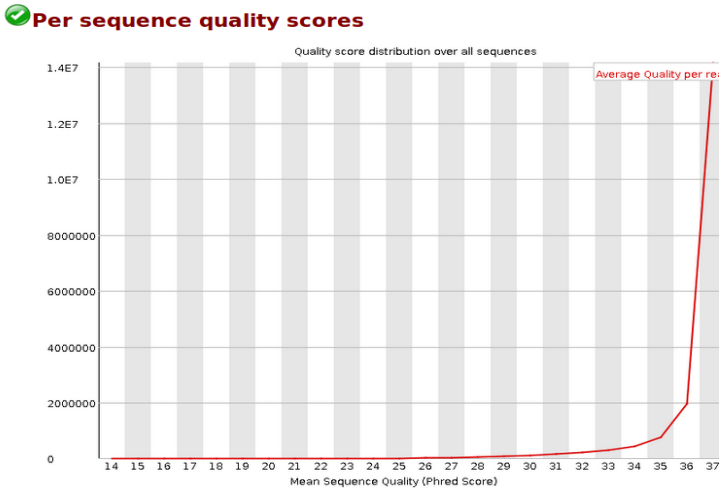


Figura 15. Valores de calidad por secuencia

Como conclusión, las gráficas FastQC de valores de calidad por secuencia muestran proporción respecto a la calidad de las secuencias por lo que no es necesario descartar ninguna de ellas.

La figura 16 da información del porcentaje de nucleótidos por cada posición en la secuencia. Líneas azules y negras por encima de rojas y verdes indicaría un %GC >%AT. Encontrar picos en las 10-13 primeras posiciones no debe de alarmarnos, es normal que exista ruido en las primeras posiciones de la secuencia RNASeq. Muchos picos distribuidos a lo largo de la secuencia indicarían un número bajo o insuficiente de secuencias.

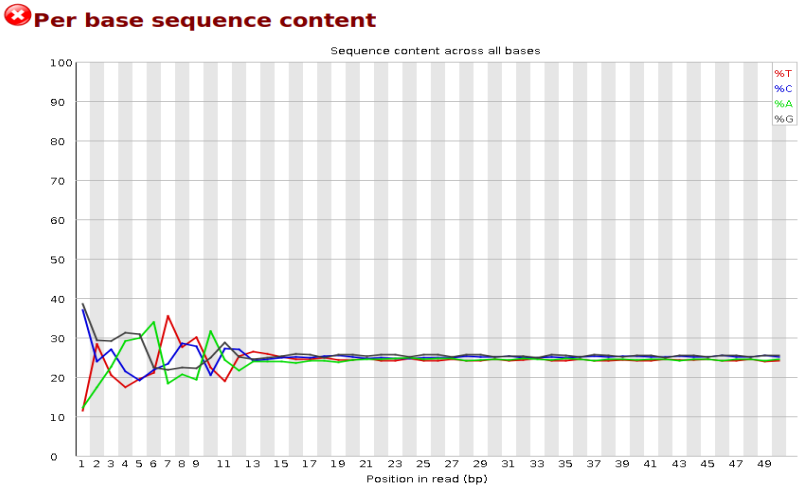


Figura 16. Contenido de secuencia por base

Como conclusión, las gráficas FastQC de calidad de secuencia por base, muestran uniformidad en el porcentaje de nucleótidos a lo largo de la secuencia, lo que expresa que el número de secuencias es suficiente.
El siguiente módulo mide el contenido GC a lo largo de la secuencia (línea roja) comparándolo con una distribución normal (línea azul). Se aprecia bastante similitud, como se puede observar en la figura 17. En caso de mala calidad puede no haber similitud para lo cual podría ser necesario eliminar el pico filtrando la secuencia.

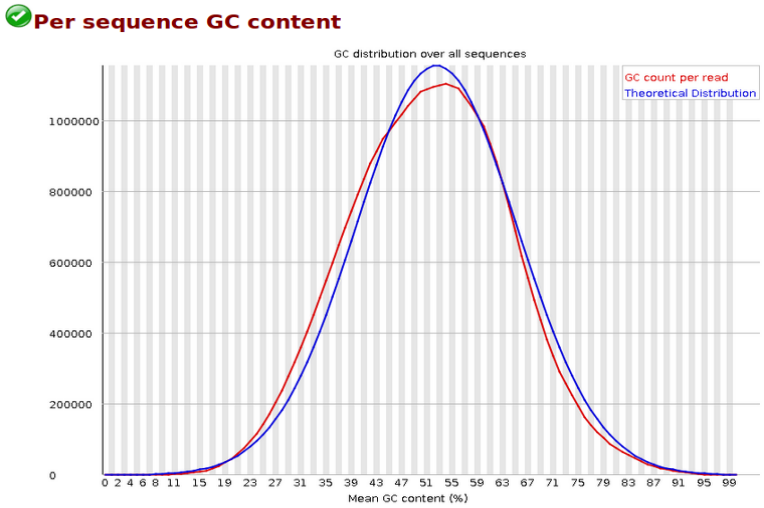


Figura 17. Contenido GC por secuencia

En las gráficas FastQC de contenido GC por secuencia, no hay picos y la distribución de las secuencias sigue una distribución normal de modo que no es necesario filtrar las secuencias.

3. Alineamiento (TopHat)

El fichero tophatAligningStatistics.xls muestra el porcentaje de alineamiento mapeados. Parte de su contenido se muestra en la figura 18. Las muestras devuelven valores en torno a un 95% lo que se considera un alto porcentaje de mapeo.

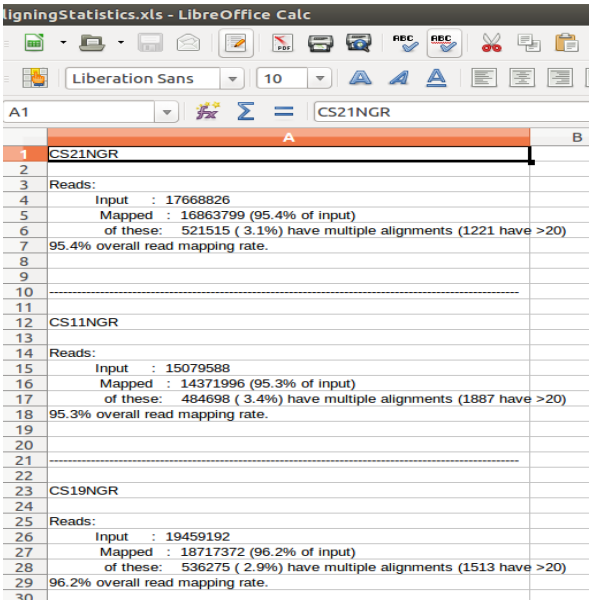


Figura 18. Porcentaje de alineamientos mapeados

Como conclusión, el conjunto de resultados FastQC y FastQScreen más el porcentaje de alineamiento de TopHat, en torno a un 95%, indica que las muestras son de buena calidad. No se considera necesario realizar el paso de trimming y downsampling debido a que el número de lecturas de las muestras no es muy dispar.

5. Expresión diferencial (cuffdiff, cuffnorm)

Se comprueba que Cuffnorm devuelve los ficheros normalizados de cada una de las comparaciones y otro para el conjunto de muestras
CS101112NGR_vs_CS123NGR.genes.fpkm_table.xls
CS192021NGR_vs_CS123NGR.genes.fpkm_table.xls
CS192021NGR_vs_CS101112NGR.genes.fpkm_table.xls
ALLsamples.genes.fpkm_table.xls

Para realizar un análisis no jerárquico, una vez revisada la calidad de las muestras, es de buenas prácticas revisar la correlación entre muestras. Ello nos da una visión global de la distribución. Se revisan las correlaciones entre muestras reportadas en el fichero ALLsamples.pearsonCorrelationsAmongSamples.xls y mostradas en la figura 19. Valores por encima de 0.9 indica que existe correlación entre las muestras.

	A	B	C	D	E	F	G	H	I	J
1		CS10NGR	CS11NGR	CS12NGR	CS19NGR	CS1NGR	CS20NGR	CS21NGR	CS2NGR	CS3NGR
2	CS10NGR	1	0.9020166124	0.9464912638	0.8625412103	0.9086863872	0.8887283603	0.8843164463	0.8655697488	0.917023381
3	CS11NGR	0.9020166124	1	0.9838105334	0.7607606325	0.8597480058	0.7946196675	0.7840180169	0.8217842719	0.8265643506
4	CS12NGR	0.9464912638	0.9838105334	1	0.8089178829	0.8963473301	0.8388747357	0.8339846048	0.8623439309	0.8749000525
5	CS19NGR	0.8625412103	0.7607606325	0.8089178829	1	0.8745844219	0.9829051947	0.9871551978	0.796914486	0.8817626053
6	CS1NGR	0.9086863872	0.8597480058	0.8963473301	0.8745844219	1	0.8985512733	0.8907387697	0.9482374059	0.9830572646
7	CS20NGR	0.8887283603	0.7946196675	0.8388747357	0.9829051947	0.8985512733	1	0.9967058579	0.8478495596	0.910721851
8	CS21NGR	0.8843164463	0.7840180169	0.8339846048	0.9871551978	0.8907387697	0.9967058579	1	0.8359086059	0.9036657239
9	CS2NGR	0.8655697488	0.8217842719	0.8623439309	0.796914486	0.9482374059	0.8478495596	0.8359086059	1	0.9555666721
10	CS3NGR	0.917023381	0.8265643506	0.8749000525	0.8817626053	0.9830572646	0.910721851	0.9036657239	0.9555666721	1
11										

Figura 19. Valores de correlación entre muestras

La figura 20 representa el mapa de correlación del conjunto de muestras. Dos componentes (PC1 y PC2) son suficientes para discriminar las muestras.

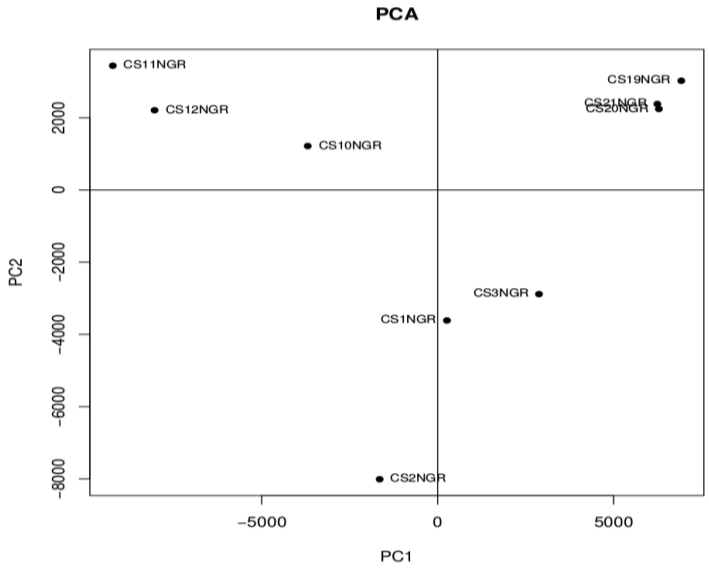
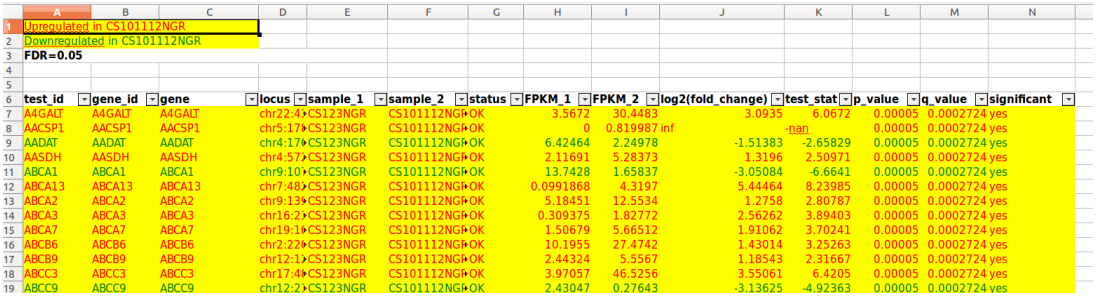


Figura 20. Mapa de correlación entre muestras

Según marcan los valores de correlación y las gráficas PCA, los valores de correlación entre las muestras indican que existe correlación con pequeñas y ligeras diferencias. La distribución de las muestras en el mapa de correlación, indica una clasificación por líneas celulares y replicados.

Se verifican los ficheros de resultados de Cuffdiff de expresión diferencial de cada una de las comparaciones realizadas.
CS101112NGR_vs_CS123NGR.gene_exp.xlsx
CS192021NGR_vs_CS123NGR.gene_exp.xlsx
CS192021NGR_vs_CS101112NGR.gene_exp.xlsx

De los resultados anteriores, Nextpresso genera nuevos ficheros filtrados por columna status = OK y un valor FPKMthreshold definido en el fichero de configuración de parámetros de Nextpresso. En nuestro caso el valor de corte es 2, por tanto, se consideran genes diferencialmente expresados aquellos cuya columna status = OK y al menos uno de los valores FPKM de las muestras sea > 2. La figura 21 corresponde a parte del fichero de resultados de expresión diferencial de la comparación entre las líneas U373 y U87. Los genes upregulados que aparecen en color rojo indican estar sobreexpresados en la muestra 2 correspondiente a la línea celular U373, mientras que los genes downregulados que aparecen en color verde, son genes diferencialmente expresados en menor cantidad en la muestra 2.
CS101112NGR_vs_CS123NGR.gene_exp_FILTERED.xlsx
CS192021NGR_vs_CS123NGR.gene_exp_FILTERED.xlsx
CS192021NGR_vs_CS101112NGR.gene_exp_FILTERED.xlsx



Upregulated in CS101112NGR
Downregulated in CS101112NGR
FDR=0.05

test_id	gene_id	gene	locus	sample_1	sample_2	status	FPKM_1	FPKM_2	log2(fold_change)	test_stat	p_value	q_value	significant
A4GALT	A4GALT	A4GALT	chr22:4	CS123NGR	CS101112NGR	OK	3.5672	30.4483	3.0935	6.0672	0.00005	0.0002724	yes
AACSP1	AACSP1	AACSP1	chr5:17	CS123NGR	CS101112NGR	OK	0	0.819987	inf	-nan	0.00005	0.0002724	yes
AADAT	AADAT	AADAT	chr4:17	CS123NGR	CS101112NGR	OK	6.42464	2.24978	-1.51383	-2.65829	0.00005	0.0002724	yes
AASDH	AASDH	AASDH	chr4:57	CS123NGR	CS101112NGR	OK	2.11691	5.28373	1.3196	2.50971	0.00005	0.0002724	yes
ABCA1	ABCA1	ABCA1	chr9:10	CS123NGR	CS101112NGR	OK	13.7428	1.65837	-3.05084	-6.6641	0.00005	0.0002724	yes
ABCA13	ABCA13	ABCA13	chr7:48	CS123NGR	CS101112NGR	OK	0.0991868	4.3197	5.44464	8.23985	0.00005	0.0002724	yes
ABCA2	ABCA2	ABCA2	chr9:13	CS123NGR	CS101112NGR	OK	5.18451	12.5534	1.2758	2.80787	0.00005	0.0002724	yes
ABCA3	ABCA3	ABCA3	chr16:2	CS123NGR	CS101112NGR	OK	0.309375	1.82772	2.56262	3.89403	0.00005	0.0002724	yes
ABCA7	ABCA7	ABCA7	chr19:1	CS123NGR	CS101112NGR	OK	1.50679	5.66512	1.91062	3.70241	0.00005	0.0002724	yes
ABCB6	ABCB6	ABCB6	chr2:22	CS123NGR	CS101112NGR	OK	10.1955	27.4742	1.43014	3.25263	0.00005	0.0002724	yes
ABCB9	ABCB9	ABCB9	chr12:1	CS123NGR	CS101112NGR	OK	2.44324	5.5567	1.18543	2.31667	0.00005	0.0002724	yes
ABCC3	ABCC3	ABCC3	chr17:4	CS123NGR	CS101112NGR	OK	3.97057	46.5256	3.55061	6.4205	0.00005	0.0002724	yes
ABCC9	ABCC9	ABCC9	chr12:2	CS123NGR	CS101112NGR	OK	2.43047	0.27643	-3.13625	-4.92363	0.00005	0.0002724	yes

Figura 21. Contenido de fichero de expresión diferencial

La tabla 8 muestra el número total de genes que han resultado diferencialmente expresados en las muestras por cada una de las comparaciones realizadas entre ellas. Además, se indica el número de genes upregulados o downregulados diferencialmente expresados en cada una de las comparaciones.

LÍNEAS COMPARADAS	TOTAL GENES DE	TOTAL GENES UPREGULADOS	TOTAL GENES DOWNREGULADOS
U373 vs U87	4357	2291	2066
LN229 vs U373	6007	2975	3032
LN229 vs U87	4860	2246	2614

Tabla 8. Resultados de genes DE entre las líneas

De los ficheros de resultados de expresión diferencial filtrados de Cuffdiff, se toman los nombres de aquellos genes expresados diferencialmente de manera significativa por cada uno de ellos y haciendo uso de la línea de comandos de linux se mezcla con el contenido de los ficheros de expresión normalizados de Cuffnorm. Además, se añade la columna Description y una cabecera específica de forma que se generan los siguientes ficheros en formato gct. La figura 22 muestra un ejemplo de contenido de fichero en formato gct.

DEGs_CS101112NGR_vs_CS123NGR.genes.fpkm_table.gct
DEGs_CS192021NGR_vs_CS123NGR.genes.fpkm_table.gct
DEGs_CS192021NGR_vs_CS101112NGR.genes.fpkm_table.gct

Los ficheros en formato gct se visualizan mediante HeatMap (figura 23) haciendo uso de la herramienta Morpheus. El color está calculado en base a los valores máximo y mínimo por cada fila. Es posible obtener el score e cada celda situando el ratón encima.

```
#1.2
4357    6
Name    Description     CS2NGR  CS3NGR  CS1NGR  CS10NGR CS12NGR CS11NGR
A4GALT  NA      4.80145 2.695   3.20514 19.7124 28.0678 43.5646
AADAT   NA      6.4036  7.41035 5.45998 2.4785  2.39731 1.87354
AASDH   NA      2.52893 1.63351 2.18828 3.60416 5.93609 6.31095
ABCA1   NA      10.6364 10.0978 20.4941 1.61476 1.77318 1.58716
ABCA13  NA      0.11086 0.109364        0.0773364       1.39826 2.80194 8.75891
ABCA2   NA      6.45233 3.88521 5.216   4.27875 10.9868 22.3947
ABCA7   NA      2.03207 1.52258 0.965728        1.7343  5.03197 10.2291
ABCB6   NA      14.7284 7.53173 8.32644 16.397  28.6835 37.342
```

Figura 22. Contenido de fichero en formato gct

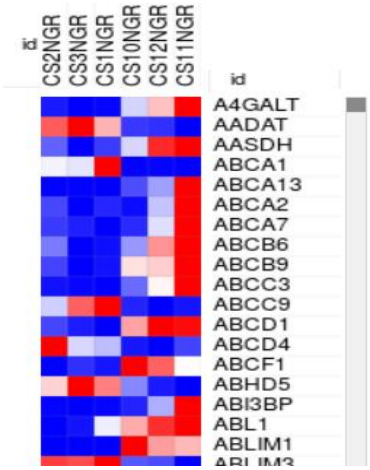


Figura 23. HeatpMap DE

Por cada comparación se realiza un seguimiento de la expresión diferencial de una lista de genes en su mayoría relacionados con angiogénesis, significativos para el investigador, teniendo en cuenta sus posibles sinónimos.

	A	B	C	D
1	ENSG	GENEID	SYNONIMOUS	GENE NAME
2	**GENES RELACIONADOS CON ANGIOGENESIS**			
3	ENSG00000112715.20.	VEGFA	VEGF, VPF	vascular endothelial growth factor A
4	ENSG00000173511.9.	VEGFB	VEGFL, VEGF	vascular endothelial growth factor B
5	ENSG00000150630.3.	VEGFC	VRP, FLT4, FLT4L	vascular endothelial growth factor C
6	ENSG00000165197.4.	VEGFD	VEGF-D, FIGF	vascular endothelial growth factor D
7	ENSG00000102755.10.	VEGFR1	FLT1, FLT, FRT	fms related tyrosine kinase 1
8	ENSG00000128052.8.	VEGFR2	KDR, CD309, FLK1, VEGFR	kinase insert domain receptor
9	ENSG00000037280.15.	VEGFR3	FLT, FLT4, PCL	fms related tyrosine kinase 4
10	ENSG00000151665.12.	PIGF	GPI11	phosphatidylinositol glycan anchor biosynthesis class F
11		CD31	PECAM1, PECA1, EndoCAM, GPIIA'	platelet and endothelial cell adhesion molecule 1
12	ENSG00000110799.13.	vWF	F8VWF	von Willebrand factor
13	ENSG00000106991.13.	CD105	ENG, END, HHT1	endoglin
14	ENSG00000100644.16.	HIF1A	bHLHe78, HIF-1alpha, HIF1, MOP1, PASD8	hypoxia inducible factor 1 alpha subunit
15	ENSG00000087245.12.	MMP2	TBE-1, CLG4A	matrix metallopeptidase 2
16	ENSG00000100985.7.	MMP9	CLG4B, GELB	matrix metallopeptidase 9
17		PDGF	HASPP28, PAP, PAP1, PDAP1	PDGFA associated protein 1
18	ENSG00000073756.11.	COX2	PTGS2, PHS II, PGHS-2, PGHS	prostaglandin-endoperoxide synthase 2
19	ENSG00000167772.11.	ANGPTL4	PGAR, PP1158, PPAR, PPARG, HFARP, NL2, FIAF, ARP4, PSEC0166, UNQ171/PRO197	angiopoietin like 4
20	ENSG00000143590.13.	EFNA3	EHK1, EHK1-L, LERK3, EPLG3, EFL-2, EPH, LERK-3, EFL2, LERK3	ephrin A3
21				
22	**OTROS GENES REVISADOS**			
23	ENSG00000141510.16.	TP53	P53, LFS1, NY-CO-13	tumor protein p53
24				
25	ENSG00000138448.11.	ITGAV	MSK8, VNRA, VTNR, CD51	integrin subunit alpha V
26	ENSG00000142208.15.	AKT	AKT, PKB, PRKBA, RAC, RAC-PK-alpha	AKT serine/threonine kinase 1
27	ENSG00000171862.9.	PTEN	MMAC1, PTEN1, TEP1, BZS, MHAM	phosphatase and tensin homolog
28	ENSG00000126453.9.	BCL2L12	BPR, BCL2 like 12	BCL2 like 12
29	ENSG00000171791.12.	BCL2	Bcl-2, PPP1R50	BCL2, apoptosis regulator
30	ENSG00000135679.21.	MDM2	HDM2, MGC5370	MDM2 proto-oncogene
31	ENSG00000149311.17.	ATM	TEL1, TELO1, ATC, ATD, ATDC, ATA	ATM serine/threonine kinase
32				
33	ENSG00000136997.15.	MYC	BHLHE39, c-Myc, MYCC, p64	MYC proto-oncogene, bHLH transcription factor
34				

Figura 24. Selección de genes para seguimiento

Finalmente, se presta especial atención a los genes VEGFA, CD105, TP53, HIF1A e ITGAV. Las figuras 25, 26 y 27 muestran el resultado de la expresión diferencial de estos genes en cada una de las comparaciones realizadas entre las líneas.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**COMPARACION 101112vs123**	**U373vsU87**												
2														
3	Upregulated in CS101112NGR													
4	Downregulated in CS101112NGR													
5	**FDR=0.05**													
6														
7														
8	**test_id**	**gene_id**	**gene**	**locus**	**sample_1**	**sample_2**	**status**	**FPKM_1**	**FPKM_2**	**log2(fold**	**test_stat**	**p_value**	**q_value**	**significan**
9	ENG (CD105)	ENG	ENG	chr9:1305	CS123NGR	CS101112NGR	OK	29.2812	15.9196	-0.87917	-2.07636	0.0006	0.002535	yes
10	ITGAV	ITGAV	ITGAV	chr2:1874	CS123NGR	CS101112NGR	OK	11.0146	25.5556	1.21423	2.93531	0.00005	0.000272	yes
11	TP53	TP53	TP53	chr17:757	CS123NGR	CS101112NGR	OK	4.42075	53.439	3.59553	5.024	0.00005	0.000272	yes
12	HIF1A	HIF1A	HIF1A	chr14:621	CS123NGR	CS101112NGR	OK	50.5989	142.183	1.49057	2.36939	0.0003	0.001375	yes
13	PTEN	PTEN	PTEN	chr10:896	CS123NGR	CS101112NGR	OK	13.0229	5.26264	-1.30719	-3.07368	0.00005	0.000272	yes
14	ATM	ATM	ATM	chr11:108	CS123NGR	CS101112NGR	OK	2.03002	3.11976	0.619943	1.39886	0.01545	0.041275	yes
15	AKT1	AKT1	AKT1	chr14:105	CS123NGR	CS101112NGR	OK	113.197	113.991	0.010079	0.023982	0.9663	0.977754	no
16	BCL2L12	BCL2L12	BCL2L12	chr19:501	CS123NGR	CS101112NGR	OK	22.1151	28.1216	0.346649	0.519246	0.3643	0.50352	no
17	MDM2	MDM2	MDM2	chr12:692	CS123NGR	CS101112NGR	OK	6.72916	6.18679	-0.12123	-0.28618	0.6182	0.7314	no
18	VEGFA	VEGFA	VEGFA	chr6:4373	CS123NGR	CS101112NGR	OK	188.04	288.446	0.617258	1.16742	0.04655	0.103471	no
19	BCL2	BCL2	BCL2	chr18:607	CS123NGR	CS101112NGR	NOTEST	0.458197	0.554757	0.275886	0	1	1	no

Figura 25. Comparación DE CS101112NGR_vs_CS123NGR.gene_exp

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**COMPARACION 192021vs123**	**LN229vsU87**												
2														
3	Upregulated in CS192021NGR													
4	Downregulated in CS192021NGR													
5	**FDR=0.05**													
6														
7														
8	**test_id**	**gene_id**	**gene**	**locus**	**sample_1**	**sample_2**	**status**	**FPKM_1**	**FPKM_2**	**log2(fold**	**test_stat**	**p_value**	**q_value**	**significant**
9	VEGFA	VEGFA	VEGFA	chr6:4373	CS123NGR	CS192021NGR	OK	201.271	37.1879	-2.43624	-6.49934	0.00005	0.000184	yes
10	HIF1A	HIF1A	HIF1A	chr14:621	CS123NGR	CS192021NGR	OK	54.1645	113.962	1.07313	2.1045	0.00155	0.004205	yes
11	ENG (CD105)	ENG	ENG	chr9:1305	CS123NGR	CS192021NGR	OK	31.3358	62.7404	1.00159	3.0194	0.00005	0.000184	yes
12	TP53	TP53	TP53	chr17:757	CS123NGR	CS192021NGR	OK	4.73058	31.1761	2.72035	4.61653	0.00005	0.000184	yes
13	ITGAV	ITGAV	ITGAV	chr2:1874	CS123NGR	CS192021NGR	OK	11.7895	30.2093	1.35749	3.94831	0.00005	0.000184	yes
14	PTEN	PTEN	PTEN	chr10:896	CS123NGR	CS192021NGR	OK	13.9374	10.1036	-0.46409	-1.36053	0.0193	0.039233	yes
15	BCL2	BCL2	BCL2	chr18:607	CS123NGR	CS192021NGR	OK	0.490275	2.17463	2.14911	2.38129	0.00035	0.001098	yes
16	AKT1	AKT1	AKT1	chr14:105	CS123NGR	CS192021NGR	OK	121.146	69.9062	-0.79325	-2.34809	0.00005	0.000184	yes
17	MDM2	MDM2	MDM2	chr12:692	CS123NGR	CS192021NGR	OK	7.20193	19.5224	1.43868	4.14592	0.00005	0.000184	yes
18	BCL2L12	BCL2L12	BCL2L12	chr19:501	CS123NGR	CS192021NGR	OK	23.6674	17.317	-0.45071	-0.79208	0.16475	0.242965	no
19	ATM	ATM	ATM	chr11:108	CS123NGR	CS192021NGR	OK	2.17239	2.17192	-0.00032	-0.00085	0.9982	0.998451	no

Figura 26. CS192021NGR_vs_CS123NGR.gene_exp

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**COMPARACION 192021vs101112**	**LN229vsU373**												
2														
3	Upregulated in CS192021NGR													
4	Downregulated in CS192021NGR													
5	**FDR=0.05**													
6														
7														
8	**test_id**	**gene_id**	**gene**	**locus**	**sample_1**	**sample_2**	**status**	**FPKM_1**	**FPKM_2**	**log2(fold**	**test_stat**	**p_value**	**q_value**	**significa**
9	VEGFA	VEGFA	VEGFA	chr6:4373	CS101112	CS192021NGR	OK	289.044	34.3933	-3.07109	-7.50613	0.00005	0.000244	yes
10	ENG (CD105)	ENG	ENG	chr9:1305	CS101112	CS192021NGR	OK	15.9459	58.0307	1.86363	5.10761	0.00005	0.000244	yes
11	TP53	TP53	TP53	chr17:757	CS101112	CS192021NGR	OK	53.6163	28.8344	-0.89488	-2.13407	0.0002	0.000868	yes
12	PTEN	PTEN	PTEN	chr10:896	CS101112	CS192021NGR	OK	5.28346	9.34486	0.82269	2.2465	0.0001	0.000464	yes
13	BCL2L12	BCL2L12	BCL2L12	chr19:501	CS101112	CS192021NGR	OK	28.2985	16.0165	-0.82117	-1.61014	0.00545	0.015469	yes
14	BCL2	BCL2	BCL2	chr18:607	CS101112	CS192021NGR	OK	0.557401	2.01145	1.85145	2.69877	0.00005	0.000244	yes
15	AKT1	AKT1	AKT1	chr14:105	CS101112	CS192021NGR	OK	114.509	64.6563	-0.8246	-2.3128	0.00005	0.000244	yes
16	MDM2	MDM2	MDM2	chr12:692	CS101112	CS192021NGR	OK	6.21998	18.0559	1.53749	4.34857	0.00005	0.000244	yes
17	ATM	ATM	ATM	chr11:108	CS101112	CS192021NGR	OK	3.12714	2.0088	-0.63851	-1.69438	0.0031	0.009598	yes
18	HIF1A	HIF1A	HIF1A	chr14:621	CS101112	CS192021NGR	OK	142.649	105.404	-0.43653	-0.92231	0.1102	0.195797	no
19	ITGAV	ITGAV	ITGAV	chr2:1874	CS101112	CS192021NGR	OK	25.6425	27.9407	0.123831	0.348238	0.5471	0.668272	no

Figura 27. CS192021NGR_vs_CS101112NGR.gene_exp

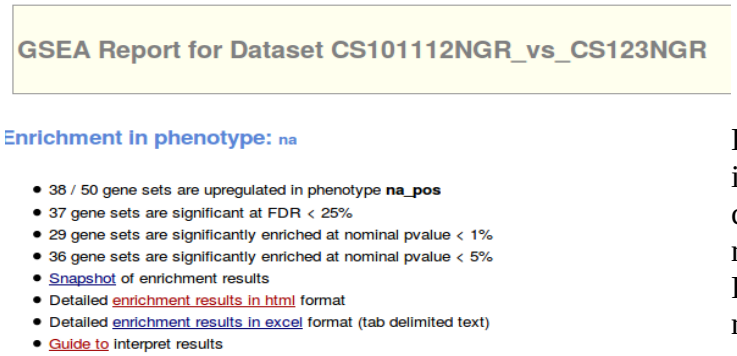
La tabla 9 representa la comparación de cantidad de expresión diferencial de los genes en cada una de las comparaciones de muestras realizadas. Se ha de tener en cuenta que las comparaciones se realizan exclusivamente entre dos elementos por lo que no es posible sacar conclusiones a tres elementos ya que la normalización se realiza entre las muestras que participan en la comparación exclusivamente.

GEN\COMPARACIÓN	U373vsU87	LN229vsU87	LN229vsU373
VEGFA		LN229<U87	LN229<U373
ENG(CD105)	U373<U87	LN229>U87	LN229>U373
HIF1A	U373>U87	LN229>U87	LN229>U373
TP53	U373>U87	LN229>U87	LN229<U373
ITGAV	U373>U87	LN229>U87	

Tabla 9. Comparativa de DE

8. Análisis funcional GSEA

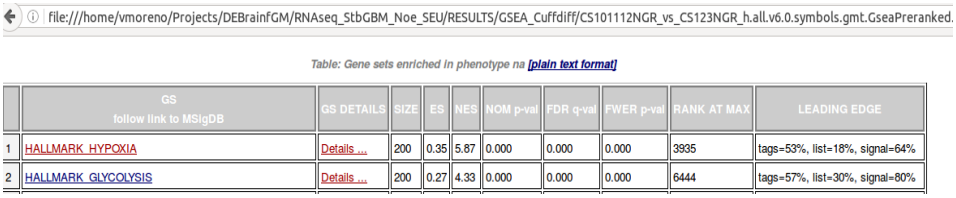
Nextpresso devuelve en el directorio GSEA_Cuffdiff los resultados en una carpeta por cada colección revisada de cada una de las comparaciones realizadas entre las muestras. Se encuentran diferentes rutas activas.
La figura 28 muestra el reporte de uno de los resultados obtenidos de GSEA.



El informe muestra información de los conjuntos de genes resultantes, valor de corte FDR < 25%, enlace a resultados, etc...

Figura 28. Reporte GSEA

La figura 29 muestra valores de cada ruta enriquecida de la colección H para la comparación realizada CS101112NGR vs CS123NGR, siendo NES y FDR los principales valores de consulta.



	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	HALLMARK_HYPOXIA	Details ...	200	0.35	5.87	0.000	0.000	0.000	3935	tags=53%, list=18%, signal=64%
2	HALLMARK_GLYCOLYSIS	Details ...	200	0.27	4.33	0.000	0.000	0.000	6444	tags=57%, list=30%, signal=80%

Figura 29. Pathways enriquecidas

Las figuras 30 y 31 muestran respectivamente, el resumen de valores de GSEA para la ruta HYPOXIA y un gráfico con los valores Enrichment Score (ES) de los genes que participan en la ruta. Los de valor positivo son genes upregulados y los de valor negativo son genes downregulados, contribuyendo en su conjunto al enriquecimiento de la ruta. En el gráfico de barras bajo la línea de valores ES, se aprecia que los genes que más contribuyen al enriquecimiento de la ruta son los que se encuentran al principio del ranking.

Table: GSEA Results Summary

Dataset	CS101112NGR_vs_CS123NGR
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	HALLMARK_HYPOXIA
Enrichment Score (ES)	0.35169458
Normalized Enrichment Score (NES)	5.865268
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

Figura 30. Resumen resultados GSEA

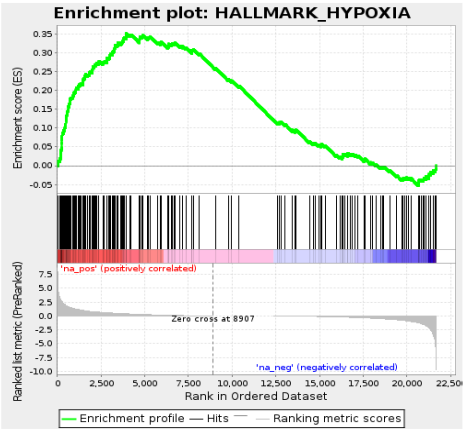


Fig 1: Enrichment plot: HALLMARK_HYPOXIA
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Figura 31. Gráfico de enriquecimiento ruta HYPOXIA

La figura 32 muestra un breve listado de genes ordenados por ranking de la ruta HYPOXIA.

Fig 1: Enrichment plot: HALLMARK_HYPOXIA
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details ***[plain text format]***

	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMEN
1	SDC3	SDC3 Entrez, Source	syndecan 3 (N-syndecan)	17	5.478	0.0042	Yes
2	PGF	PGF Entrez, Source	placental growth factor, vascular endothelial growth factor-related protein	24	5.296	0.0089	Yes
3	GPC4	GPC4 Entrez, Source	glypican 4	40	4.889	0.0132	Yes
4	CP	CP Entrez, Source	ceruloplasmin (ferroxidase)	119	3.381	0.0146	Yes

Figura 32. Listado de genes ordenados por ranking de la ruta enriquecida HYPOXIA

Se revisan las pathways más significativas de cada colección y se presta especial atención a aquellas rutas relacionadas con angiogénesis.

Del hallmark h.all.v6.0.symbols de cada una de las comparaciones entre las muestras, encontramos activas las siguientes rutas, con los valores de Normalized EScore y FDR indicados:

HALLMARK_HYPOXIA
U373 vs U87 5.87 (NES) 0.00 (FDR)
LN229 vs U87 2.44 (NES) 0.001 (FDR)
Enriquecimiento de ruta en U373 > U87
Enriquecimiento de ruta en LN229 > U87

HALLMARK_P53_PATHWAY
LN229 vs U87 3.37 (NES) 0.00 (FDR)
U373 vs U87 3.00 (NES) 0.00 (FDR)
LN229 vs U373 2.95 (NES) 0.00 (FDR)
Enriquecimiento de ruta en LN229 > U87 y U373
Enriquecimiento de ruta en U373 > U87

HALLMARK_ANGIOGENESIS
LN229 vs U87 2,12 (NES) 0.004 (FDR)
U373 vs U87 1.73 (NES) 0.029 (FDR)
LN229 vs U373 1.26 (NES) 0.204 (FDR)
Enriquecimiento de ruta en LN229 > U87 y U373
Enriquecimiento de ruta en U373 > U87

HALLMARK_APOPTOSIS
LN229 vs U87 2.98 (NES) 0.00 (FDR)
LN229 vs U373 2.38 (NES) 0.001 (FDR)
U373 vs U87 2.36 (NES) 0.001 (FDR)
Enriquecimiento de ruta en LN229 > U87 y U373
Enriquecimiento de ruta en U373 > U87

La tabla 10 representa la comparación de enriquecimiento de rutas en cada una de las comparaciones realizadas entre muestras. Se ha de tener en cuenta que las comparaciones se realizan exclusivamente entre dos elementos por lo que no es posible sacar conclusiones a tres elementos ya que los valores normalizados se obtienen exclusivamente entre las muestras participantes.

RUTA\COMPARACIÓN	U373vsU87	LN229vsU87	LN229vsU373
H_HYPOXIA	U373>U87	LN229>U87	
H_ANGIOGÉNESIS	U373>U87	LN229>U87	LN229>U373
H_APOPTOSIS	U373>U87	LN229>U87	LN229>U373
H_P53	U373>U87	LN229>U87	LN229>U373

Tabla 10. Comparativa de enriquecimiento de rutas

9. Predicción de fusión de genes (TopHat-Fusion)

No se encuentran genes fusión en el resultado.

7. DISCUSIÓN

7.1 DISCUSIÓN DEL SISTEMA DE INFORMACIÓN

Lo primero que se discute es la elección de ciertas características del sistema como por ejemplo la elección del sistema gestor de base de datos, el lenguaje apropiado para el desarrollo de la aplicación web, y algunos detalles específicos que a continuación se comentan:

- Dados los requerimientos de nuestra base de datos, acceso vía web, diseño no demasiado complejo, necesidad de representación de datos específicos no contemplados en SQLite y la importancia de la velocidad de acceso a los datos, se considera como opción más apropiada el uso del RDBMS MySql.

- Se recomienda el uso de las herramientas Django y Python para el desarrollo de aplicaciones por ser un lenguaje sencillo, fácil de usar, familiarizado con el entorno de Linux (entorno de trabajo habitual en biomedicina), que permite amplia funcionalidad y consulta de librerías específicas en el área de la bioinformática.

- El motivo de que la gran mayoría de campos de la entidad RESULTDE (entidad que alberga datos de detalles de resultados) del modelo de datos sean opcionales, es no interferir en los procesos de carga masiva ya que diferentes resultados de detalle de análisis por técnicas o empresas distintas puede aportar distinta información y contenido. Tratándose de campos opcionales, el sistema no se vería perjudicado en caso de no cargar alguno de dichos valores. De otra manera, sólo se podrían cargar resultados que tuviesen al menos la información requerida en nuestro sistema.

Como es normal, un sistema de información no lo abarca todo por lo que se encuentran limitaciones del propio sistema que se desarrollan en los siguientes puntos:

- El proceso de carga masiva de datos requiere de un proceso independiente y deberá seguir un modelo evolutivo para adaptarse a los diferentes formatos en los que se reciba la información. Puede llegar información en ficheros ordenados por columnas o en formato texto separado por un determinado carácter, pueden no corresponderse las columnas con los campos de la base de datos, etc... Por este motivo, se podría llegar a necesitar el uso de procesos automáticos independientes para realizar un preprocesado de la información previa carga de datos.

- En el apartado de resultados, especificaciones del sistema de información se propuso el hecho de que la aplicación web mostrase un gráfico tipo boxplot ordenando las líneas celulares de mayor a menor expresión de un gen dado por el investigador. Evaluado este propósito, se considera inviable dada la información disponible en nuestro sistema ya que se almacenan valores de expresión diferencial comparativos entre líneas celulares y no valores de expresión individual por línea. Tampoco es posible obtener esta información mediante cálculos porque necesitaríamos datos no almacenados en el sistema. Por tanto, se propone mostrar un barplot ordenando por comparaciones de líneas

celulares, de mayor a menor expresión de un gen dado por el investigador o almacenar en el sistema de información los valores de expresión por línea celular.

- Los resultados de un análisis de RNASeq son muchos y muy variados, se ha realizado una selección de modo que el modelo de datos se ha diseñado para albergar principalmente detalles de resultados de expresión diferencial de la comparación entre dos líneas celulares. Una ventaja es que el diseño facilita la inserción de nuevos tipos de detalles de resultados. Para almacenar otro tipo de resultados que no sean de DE de RNASeq se deberán crear las entidades específicas, así como sus atributos correspondientes.

- Si se desea por alguna circunstancia almacenar información de expresión de una línea celular, lo ideal como ya se ha mencionado, sería crear una nueva entidad con los atributos correspondientes. Otra opción menos recomendada pero sí útil sería almacenar esos resultados en la tabla RESULTDE habiendo marcado como opcionales los campos relativos a muestra 2.

A pesar de las limitaciones mencionadas el sistema de información cuenta con una serie de ventajas principales como son:

- Permite una manipulación rápida y de fácil acceso a la información.
- Organiza y almacena la información de investigación relevante, de modo que esté disponible cuando sea necesaria para futuros análisis.
- Ofrece, de forma transparente al usuario, la posibilidad de realizar procesos de tratamiento automático de la información que permitan dar respuesta a preguntas complejas imposibles de responder de manera manual.

7.2 DISCUSIÓN DE RESULTADOS DEL ANÁLISIS DE EXPRESIÓN DIFERENCIAL

Analizadas las características de RNASeq frente a otros métodos transcriptómicos y a pesar de ser un método con gran carga computacional se considera la técnica más apropiada para la secuenciación de nuestras líneas, como técnica de segunda generación NGS (Next Generation Sequencing), dando especial relevancia a la resolución que esta técnica ofrece. (X Wang, Z. et al. 2009)

El análisis de expresión diferencial se realiza con la herramienta Nexpresso, un pipeline que integra distintas herramientas y permite llevar a cabo un análisis completo de los datos de RNASeq. Se ejecuta por docker permitiendo así una mayor abstracción, manejabilidad y facilidad de uso de la herramienta. (Graña, O. et al. 2017).

Comentar que el uso de Cuffnorm para la cuantificación de las muestras permite una normalización de los valores más ajustada debido a su cálculo intra e inter muestra. Por este motivo no se hace uso de Cufflinks. Como se han obtenido resultados de Cuffnorm, no se atienden a los resultados de cuantificación y expresión diferencial de Htseq-count y DESeq2**.**

Como era de esperar, en el estudio de fusión de genes mediante TopHat-fusion (Kim, D. y Salzberg, S. L. 2011), no tenemos resultados en nuestro experimento Single End ya que cobra sentido realizarlo cuando se trata de análisis Paired End en los que se secuencian los dos extremos de la molécula.

Para entender los resultados del análisis RNASeq DE y GSEA en nuestro estudio, es importante situarnos en el contexto de neoangiogénesis en el que cobran vida los genes VEGF, HIF1A y el marcador de endotelio CD105. Se introducen estos conceptos.

VEGF se ha encontrado regulado en más del 60% de los casos de GBM como un biomarcador y objetivo ideal (McNamara, M. G. et al. 2013).

VEGF es una proteína de señalización regulada por una tirosina quinasa, que desempeña un papel clave en la angiogénesis, y la permeabilidad de la barrera hematoencefálica (BBB) (Feng, S. y Huan, Y. 2011).

VEGF se ha descubierto en células GBM, con una mayor expresión en torno a áreas de tejido necrótico; cuando el tumor tiene áreas de hipoxia, las células circundantes producen un factor inducible por hipoxia (HIF) que estimula la liberación de VEGF. VEGF se une a los receptores de VEGF (VEGFR) que están presentes en las superficies de las células; activar un bucle autocrino presente causa transfosforilación y esto conduce a la angiogénesis (Clara, C. A. et al. 2014; Krcek, R. et al. 2017; Takata, K. et al. 2008; Turkowski, K. et al. 2018)

La angiogénesis es esencial para el crecimiento y la metástasis de neoplasias malignas sólidas. El recuento de vasos tumorales y la expresión del factor de crecimiento endotelial vascular (VEGF), un potente factor angiogénico, se han asociado con el pronóstico (Moghaddam, N.A. et al. 2015)

La figura 33 muestra un esquema simple del proceso por el cual la expresión de la proteína VEGFA induce la proliferación de vasos sanguíneos.

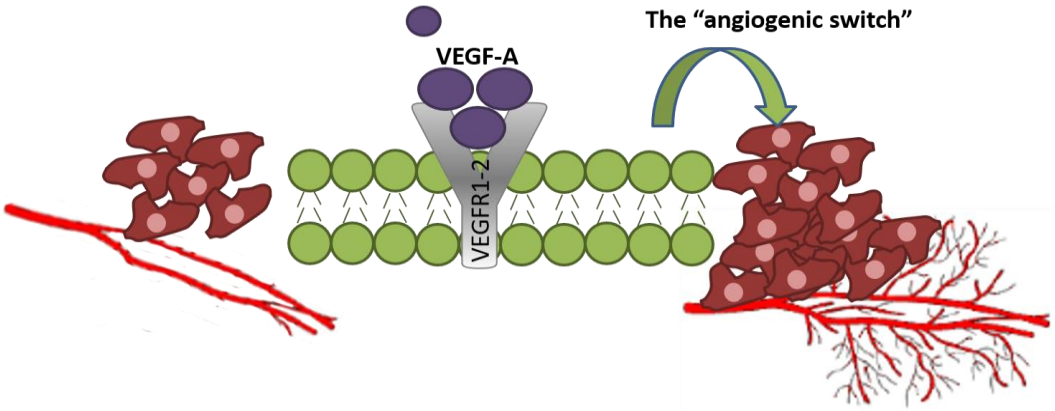


Figura 33. Proliferación de vasos sanguíneos por expresión de la proteína VEGFA (García-Romero et al. 2018)

Hay cuatro tipos de VEGF en el cuerpo humano, VEGFA-D, con VEGFA como la proteína de señalización dominante con más subtipos (ver figura 34). Estos cuatro

factores de crecimiento se unen a los diferentes VEGFR, VEGFR1-3, con VEGFR2 como el receptor dominante (Randi, A.M. et al. 2013). Se cree que VEGFR1 modula la señalización de VEGFR2. VEGFA se une a VEGFR1 y 2 y, a través de la interrupción de esta vía, se ha demostrado que detiene la proliferación de las células tumorales.

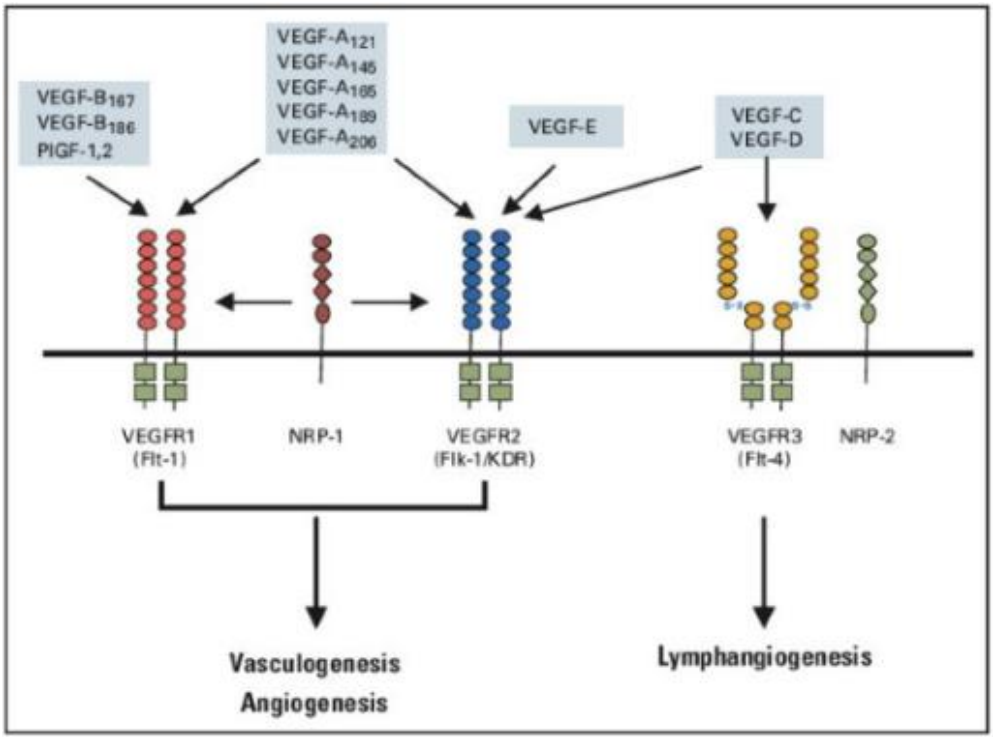


Figura 34. Diagrama que explica dónde se une VEGFA y su función principal al receptor. Su bloqueo puede conducir a la disminución de angiogénesis a través de VEGFR2 o mediante la unión a VEGFA (Katzel, J.A. et al. 2009)

Estudios previos revelan que la endoglina (ENG) o marcador de endotelio CD105 es un marcador de angiogénesis sensible en tejidos neoplásicos. (Moghaddam, N. A. et al 2015; Sica, G. et al. 2011)

La endoglina o marcador de endotelio CD105 (grupo de moléculas de diferenciación 105) es una proteína homodímera transmembrana de 180 kDa que se encuentra característicamente en las paredes endoteliales (vasos sanguíneos). Es un componente clave de la vía de señalización del receptor del factor de crecimiento transformante b (TGFb), involucrado con los receptores 1 y 2. CD105 cumple funciones clave en angiogénesis y vasculogénesis durante el desarrollo (posiblemente mediante la prevención de la apoptosis en células endoteliales hipóxicas. Como marcador de vasos sanguíneos, CD105 se asocia con vasos inmaduros, y algunos estudios sugieren que podría ser un marcador preferentemente de nuevos vasos angiogénicos. También se ha propuesto CD105 como marcador de células madre mesenquimales. Debido a su aparente especificidad por los vasos sanguíneos asociados a tumores, CD105 también tiene interés como objetivo terapéutico, con la terapia con anticuerpos monoclonales en los ensayos clínicos en etapas tempranas. (Moghaddam, N.A. et al. 2015)

Basándonos en estudios previos del grupo liderado por el Doctor Ángel Ayuso Sacido, en los que se observaban diferencias entre las tres líneas establecidas de glioma de alto grado (U87, U373 y LN229) en cuanto a marcadores relacionados con la angiogénesis, sumaremos a nuestro estudio los resultados de algunas técnicas empleadas por este equipo, como ELISA (Enzyme-Linked ImmunoSorbent Assay, conocida como ensayo por inmunoabsorción ligado a enzimas. Se trata de una técnica de laboratorio que permite detectar pequeños segmentos de proteína) e inmunofluorescencia (técnica que se utiliza para la detección de estructuras subcelulares que nos permiten estudiarlas con la utilización de anticuerpos acoplados a fluoróforos). De esta forma podremos comparar si existe relación alguna entre tales resultados y nuestro análisis de RNASeq.
Tras el cálculo de la secreción basal de VEGFA mediante ELISA en nuestras líneas, en estado de normoxia, representado en la fitura 35, obtuvieron que la proteína VEGF se expresaba mayoritariamente en la línea U87, seguida de la línea U373 y finalmente con menor expresión en la línea LN229:

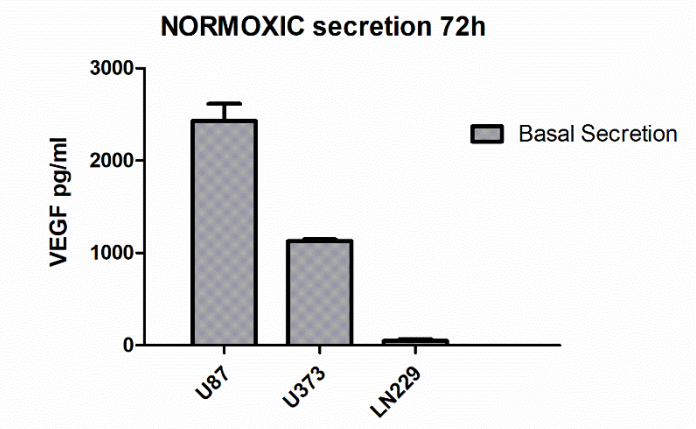


Figura 35. Resultado ELISA de expresión de VEGF

Los resultados del análisis DE de RNASeq de la comparación realizada entre las tres líneas celulares U373, U87 y LN229, en estado de normoxia, indican que el nivel del gen VEGFA expresado en la línea celular LN229 es menor que el expresado en las líneas celulares U87 y U373. Este resultado concuerda con el resultado del estudio ELISA realizado sobre las mismas líneas celulares en estado de normoxia, lo que no revela el estudio DE de RNASeq es diferencias de expresión génica significativas entre las líneas U373 y U87.

A continuación, se muestran los resultados que obtuvieron por inmunofluorescencia de las células en estudio. En la figura 36 se observa que la expresión del marcador de endotelio CD105 y la formación de vasos sanguíneos es mayor en la línea U87, seguida de la línea U373 en la que se aparecen en menor medida formación de vasos sanguíneos y por último la línea LN229 con una menor expresión y clara disminución de angiogénesis (García-Romero et al. 2018).

Los colores de las imágenes representan:
ROJO → CD105 (marcador de endotelio), neoangiogénesis
VERDE → marcador vimentina
AZUL → núcleo de las células tumorales

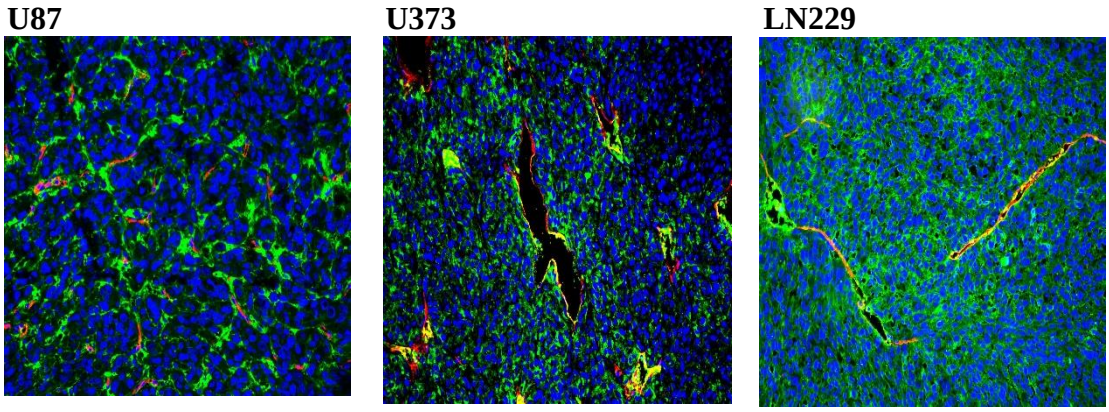


Figura 36. Imágenes de las líneas celulares U87, U373 y LN229 en inmunofluorescencia

Los resultados del análisis DE de RNASeq de la comparación realizada entre las tres líneas celulares U373, U87 y LN229, en estado de normoxia, indican que el nivel del marcador de endotelio CD105 (ENG) expresado en la línea celular LN229 es mayor que el expresado en las líneas celulares U87 y U373, a su vez, el nivel de expresión del marcador en la línea celular U373 es menor que el expresado en la línea U87.

Que el resultado del análisis DE de RNASeq respecto al marcador CD105 indique que está más expresado en la línea U87 que en la línea U373, coincide con el resultado de la inmunofluorescencia realizada sobre las mismas líneas en estado de normoxia, sin embargo, no es coincidente la expresión del marcador en la línea LN229 frente a las líneas U373 y U87.

Llegados a este punto, es interesante recordar que el análisis de RNASeq se produce a nivel de transcriptoma mientras que las técnica ELISA e inmunofluorescencia se realizan a nivel de proteoma y por tanto podrían existir diferencias en su expresión.

Si confrontamos los resultados GSEA del análisis de RNASeq con las imágenes obtenidas de la inmunofluorescencia, coincide que las rutas de angiogénesis, apoptosis, hipoxia y p53 se encuentran más activas en las líneas con menor desarrollo de vasos sanguíneos. Esto coincide con el hecho de que una menor vascularización sugiere menos oxígeno y por tanto mayor hipoxia lo que puede inducir mayor muerte celular. Además, la ruta P53 puede mediar en situaciones de apoptosis (Vazquez, A. et al 2008).

Como se ha mencionado anteriormente, si recordamos que, en las áreas de hipoxia, las células circundantes producen el factor HIF (Takata, K. 2008). El resultado DE del análisis de RNASeq realizado entre las líneas, coincide con dicha anotación e indica que el nivel de expresión del factor HIF1A es mayor en las líneas con menor vascularización, es decir, donde se encuentran áreas activas por hipoxia.

El proyecto realizado abre la puerta a nuevas vías de estudio experimental y biocomputacional.

7.3 FUTURAS LÍNEAS DE ESTUDIO

Experimental

- Realizar un análisis RNASeq de líneas tumorales establecidas U87, U373 y LN229 vs datos de líneas sanas secuenciadas del mismo modo, para establecer diferencias entre ellas y detectar posibles biomarcadores.
- Realizar un análisis RNASeq de líneas Cancer Stem Cells vs líneas sanas secuenciadas del mismo modo, para establecer diferencias entre ellas y detectar posibles biomarcadores.
- Realizar un análisis RNASeq de líneas establecidas en estado de normoxia vs estado de hipoxia secuenciadas del mismo modo, para establecer diferencias entre ellas y detectar posibles biomarcadores.
- Realizar un análisis RNASeq de líneas Cancer Stem Cells en estado de normoxia vs estado de hipoxia secuenciadas del mismo modo, para establecer diferencias entre ellas y detectar posibles biomarcadores.
- Analizar variantes de splicing y SNPs en líneas Cancer Stem Cells de glioma para detectar posibles biomarcadores.
- Estudiar efectos de fármacos inhibidores de la función de angiogénesis en marcadores específicos como VEGF y CD105 para facilitar nuevos tratamientos médicos.

Biocomputacional

<u>Relativos al sistema de información de gliomas</u>

- Desarrollo de una aplicación web para el sistema de información de gliomas en base al modelo de datos diseñado y siguiendo las especificaciones técnicas descritas en el apartado de resultados.
- Desarrollo de un módulo de servidor para la carga masiva de detalles de resultados en el sistema de información de gliomas.
- Realizar diferentes tipos de gráficos como plots, boxplots, diagramas, etc. representativos de resultados del análisis RNASeq realizado para facilitar su comprensión al investigador.
- Añadir al modelo de datos resultados de análisis distintos de RNASeq como por ejemplo DNASeq. Si nos basamos en el modelo de datos realizado, este paso requiere por cada nuevo detalle de resultados crear una entidad con formato específico que representen las características que se desean almacenar del nuevo resultado.
- Crear un proceso que dado una serie de formatos RNASeq calcule la expresión de una sola línea celular. Según nuestro modelo de datos, una línea celular puede contener uno o varios formatos que cada uno de ellos se corresponde con una secuenciación. Los análisis de expresión diferencial no ofrecen datos de expresión a nivel exclusivo de línea celular ya que los cálculos se realizan intra e inter muestra. Disponer de esta información puede ser útil para realizar análisis de clustering u otro tipo de análisis diferentes de RNASeq. Tal fin requiere, como se ha comentado en el punto anterior, adaptar el modelo de datos para añadir una nueva entidad con formato específico.

Nuevos desarrollos

- Desarrollo de aplicaciones específicas que requieran manipulación de los datos del sistema de información diseñado, para cubrir futuras necesidades propuestas por el investigador.
- Realizar clustering de distintas líneas tumorales de Glioblastoma mediante el uso de técnicas de machine learning para establecer clasificaciones o análisis de grupos que puedan aportar nuevas perspectivas.
- Usar el modelo de datos para adaptar resultados concretos a herramientas de bases de datos ya diseñadas como la desarrollada por los ingenieros de la Universidad Politécnica de Madrid (UPM) que permite un tratamiento concreto de la información, visualizando resultados de manera gráfica.
- Generar una red de regulación de sistema nodal. Este tipo de redes modela la relación entre elementos de un sistema de información y diferentes estudios realizados. Siendo su objetivo principal, facilitar al usuario una visualización gráfica, difícilmente representable por otros medios, debido a la gran cantidad de información que maneja. Si además, se tiene en cuenta la teoría de grafos aplicada a redes que evolucionan en el tiempo, la red permite partir de patrones estacionarios y predecir un comportamiento futuro, de forma que, podría detectar nuevas relaciones desconocidas entre los elementos que representa. En el mercado ya existen herramientas que facilitan la generación de este tipo de redes como por ejemplo la diseñada por la empresa Next Limit.

7.4 PROBLEMAS ENCONTRADOS DURANTE LA REALIZACIÓN DEL PROYECTO

El proyecto inicial "Análisis de variantes de AS y SNPs" se enmarcaba en un proyecto de mayor ámbito, en el que posteriormente a su comienzo, se quiso realizar una base de datos para hacer uso de herramientas como la red ofrecida por Next Limit y la herramienta de visualización gráfica y análisis de datos de UPM.
Durante bastantes semanas y reuniones varias con las diferentes empresas y la FiHM, se conocieron sus herramientas, se reunió suficiente información y se generó un documento partiendo de esta idea inicial. Teniendo en cuenta que este proyecto tenía un carácter de prácticas y formación durante un tiempo limitado, fue entonces cuando llegamos a la conclusión de que dicho proyecto tenía envergadura de tesis y no de trabajo fin de máster (TFM) por lo que era inviable realizarlo. Habiendo ya avanzado bastante tiempo del periodo de prácticas se tuvo que reinventar el proyecto y se decidió realizar un análisis de expresión diferencial de RNASeq entre diferentes tipos de líneas tumorales de la FiHM (líneas establecidas y Cancer Stem Cells (CSCs)).
A medida que nos introducíamos en el proceso, comenzamos a tener problemas para reunir la información necesaria como, por ejemplo, características técnicas de la secuenciación de los formatos en RNASeq de los distintos tipos de líneas celulares, desconocida por los investigadores. El análisis de datos RNASeq se debe de realizar en función de la tecnología y métodos empleados en la secuenciación realizada. La falta de información, ya que las líneas se habían secuenciación años antes e incluso alguno de los responsables de los estudios ya no formaba parte del equipo de la FiHM, dificultaba aún más el acceso a dicha información lo que supuso varias semanas en proceso de investigación de los datos. Por este motivo se descartó utilizar las líneas CSCs y se

valoró realizar el análisis entre las tres líneas tumorales U373, U87 y LN229 de las que teníamos más información vs líneas sanas. Como no teníamos información de líneas sanas y ya no había tiempo para seguir investigando (se pueden obtener datos de líneas sanas de TCGA), finalmente, se decidió realizar el análisis de expresión diferencial entre las tres líneas tumorales.

Se aprovechó el aprendizaje de la pérdida de tiempo para reunir la información relevante previa de los formatos, para decidir diseñar un sistema de información que permitiese evitar estos problemas en un futuro y sirviera de repositorio de datos de resultados de detalles de las líneas y de sus experimentos.

Otra de las principales dificultades encontradas ha sido la comunicación y dificultad de comprensión entre las partes involucradas: investigadores, ingenieros y bioinformáticos implicando un algo coste en tiempo de reuniones y puestas en común.
El principal problema de los investigadores es la manipulación de grandes cantidades de datos (datos multiómicos) que requieren, para su tratamiento, de procesos informáticos.
El principal problema de un informático es la necesidad de comprensión del problema biológico para entender relaciones entre los datos.
En este aspecto, el papel del bioinformático es fundamental:

- utiliza un lenguaje que facilita la comprensión entre las partes
- tiene una visión genérica del problema más amplia basada en su capacitación en áreas clínicas que pueda dar sentido a los datos multiómicos (Gómez-López, G. et al. 2017)

Destacar que ser el único bioinformático en una empresa no es fácil, ya que a priori nadie comprende tu perspectiva.

Debido a mi perfil previo de ingeniera informática, una gran dificultad encontrada a lo largo del proyecto, ha sido la comprensión de información biológica muy concreta lo que ha implicado muchas horas trabajo y esfuerzo pero que finalmente ha tenido su recompensa.

A título personal, las dificultades encontradas no han sido un impedimento para realizar el trabajo de TFM, a excepción de las líneas abiertas que no ha dado tiempo a finalizar como el desarrollo de la aplicación web y el modelo físico de base de datos, sino que han formado parte de un aprendizaje continuo y real, de diferentes necesidades y colaboración entre las distintas partes, cuya finalidad ha sido proporcionar soluciones y gestionar de una manera coherente la información.

8. CONCLUSIONES

Se ha diseñado un sistema de información de gliomas para la FiHM con el propósito principal de albergar los datos de detalles de resultados de análisis de expresión diferencial de los diferentes formatos de las líneas celulares en estudio y almacenar características técnicas consideradas información relevante y necesaria para futuros usos experimentales. Por ejemplo, información de la tecnología y métodos empleados en la secuenciación de los formatos.
El modelo de datos permite a su vez tener la información actualizada, organizada y relacionada, de modo que, sea fácilmente accesible y localizable por el grupo de usuarios investigadores, facilitando así la labor de mantenimiento de la información y evitando perder tiempo en proyectos experimentales destinado a la recuperación de los datos.

En el resultado del análisis de expresión diferencial de RNASeq de las líneas celulares tumorales de GBM (U87, U373 y LN229), se han encontrado activas las rutas relacionadas con el proceso de neoangiogénesis y en su comparación con otros ensayos realizados sobre las mismas líneas, coincide una mayor expresión en las líneas celulares con menor desarrollo de vasos sanguíneos. Teniendo en cuenta que se han visto en las líneas, genes diferencialmente expresados como VEGF, CD105 y HIF1A, también relacionados con el proceso de neoangiogénesis, sugerimos nuevos estudios para un posible uso como biomarcadores.

9. REFERENCIAS BIBLIOGRÁFICAS

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., ... & Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta neuropathologica*, *114*(2), 97-109.

McNamara, M. G., Sahebjam, S., & Mason, W. P. (2013). Emerging biomarkers in glioblastoma. *Cancers*, *5*(3), 1103-1119.

Skog, J., Würdinger, T., Van Rijn, S., Meijer, D. H., Gainche, L., Curry Jr, W. T., ... & Breakefield, X. O. (2008). Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature cell biology*, *10*(12), 1470.

Bell, E. H., Hadziahmetovic, M., & Chakravarti, A. (2011). Evolvement of molecular biomarkers in targeted therapy of malignant gliomas. In *Brain Tumors-Current and Emerging Therapeutic Strategies*. InTech.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., & Walters, L. (1998). New goals for the US human genome project: 1998-2003. *science*, *282*(5389), 682-689.

Collins, F. S., & McKusick, V. A. (2001). Implications of the Human Genome Project for medical science. *Jama*, *285*(5), 540-544.

Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., & Al-Shahrour, F. (2017). Precision medicine needs pioneering clinical bioinformaticians. *Briefings in bioinformatics*.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57.

Graña, O., Rubio-Camarillo, M., Fdez-Riverola, F., Pisano, D. G., Glez-Peña, D., Nextpresso: Next Generation Sequencing Expression Analysis Pipeline, Current Bioinformatics, 2017, Volume 12, DOI: 10.2174/1574893612666170810153850

Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology*, *12*(8), R72.

Feng, S., Huang, Y., & Chen, Z. (2011). Does VEGF secreted by leukemic cells increase the permeability of blood–brain barrier by disrupting tight-junction proteins in central nervous system leukemia?. *Medical hypotheses*, *76*(5), 618-621.

Krcek, R., Latzer, P., Adamietz, I. A., Bühler, H., & Theiss, C. (2017). Influence of vascular endothelial growth factor and radiation on gap junctional intercellular communication in glioblastoma multiforme cell lines. *Neural regeneration research*, *12*(11), 1816.

Clara, C. A., Marie, S. K., Almeida, J. R. W., Wakamatsu, A., Oba-Shinjo, S. M., Uno, M., ... & Rosemberg, S. (2014). Angiogenesis and expression of PDGF-C, VEGF, CD105 and HIF-1α in human glioblastoma. *Neuropathology*, *34*(4), 343-352.

Takata, K., Morishige, K. I., Takahashi, T., Hashimoto, K., Tsutsumi, S., Yin, L., ... & Kurachi, H. (2008). Fasudil-induced hypoxia-inducible factor-1α degradation disrupts a hypoxia-driven vascular endothelial growth factor autocrine mechanism in endothelial cells. *Molecular cancer therapeutics*, *7*(6), 1551-1561.

Turkowski, K., Brandenburg, S., Mueller, A., Kremenetskaia, I., Bungert, A. D., Blank, A., ... & Vajkoczy, P. (2018). VEGF as a modulator of the innate immune response in glioblastoma. *Glia*, *66*(1), 161-174.

García-Romero *et al.* 2018 in press, art Dose-Dependent Bevacizumab Treatment in Glioblastoma Leads to Reduced Tumour Growth, Blood Vessel Development and Endothelial Cell Migration.

Randi, A. M., Laffan, M. A., & Starke, R. D. (2013). Von Willebrand factor, angiodysplasia and angiogenesis. *Mediterranean journal of hematology and infectious diseases*, *5*(1).

Katzel, J. A., Fanucchi, M. P., & Li, Z. (2009). Recent advances of novel targeted therapy in non-small cell lung cancer. *Journal of hematology & oncology*, *2*(1), 2.

Moghaddam, N. A., Mahsuni, P., & Taheri, D. (2015). Evaluation of endoglin as an angiogenesis marker in glioblastoma. *Iranian journal of pathology*, *10*(2), 89.

Sica, G., Lama, G., Anile, C., Geloso, M. C., La Torre, G., De Bonis, P., ... & Mangiola, A. (2011). Assessment of angiogenesis by CD105 and nestin expression in peritumor tissue of glioblastoma. *International journal of oncology*, *38*(1), 41-49.

Vazquez, A., Bond, E. E., Levine, A. J., & Bond, G. L. (2008). The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature reviews Drug discovery*, *7*(12), 979.

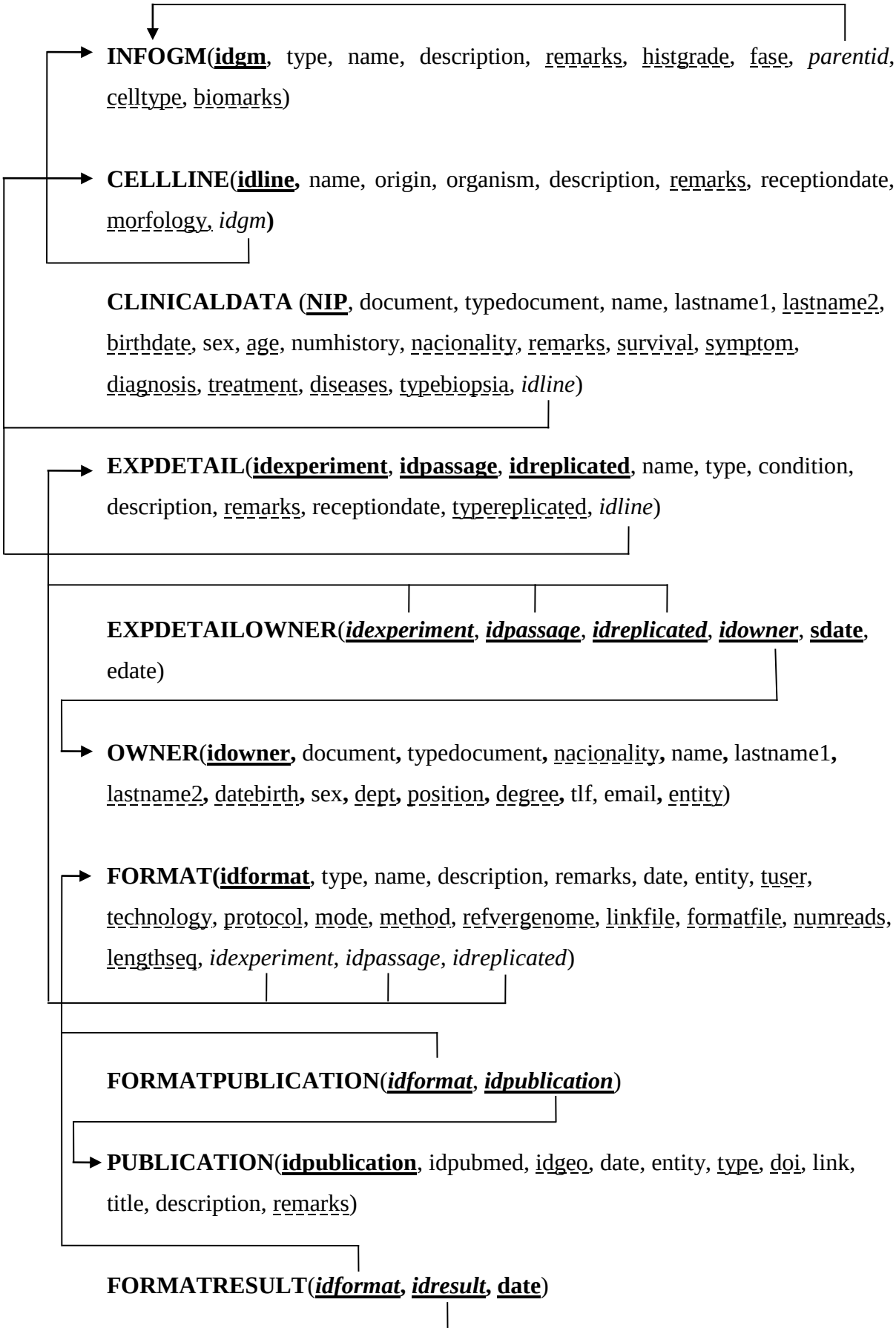
ANEXO 1. MODELO RELACIONAL Y ESPECIFICACIONES DE LA BASE DE DATOS

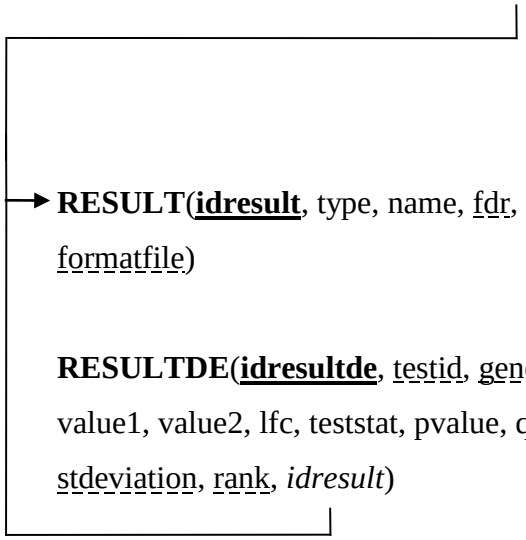
Modelo Relacional

El modelo relacional es un modelo de datos lógico que representa la transformación del diseño conceptual y su normalización para realizar un diseño físico de la base de datos.

La notación utilizada en el modelo relacional es la siguiente:

TIPO	SIGNIFICADO
negrita	Clave principal
-------	Atributo opcional
cursiva	Clave ajena

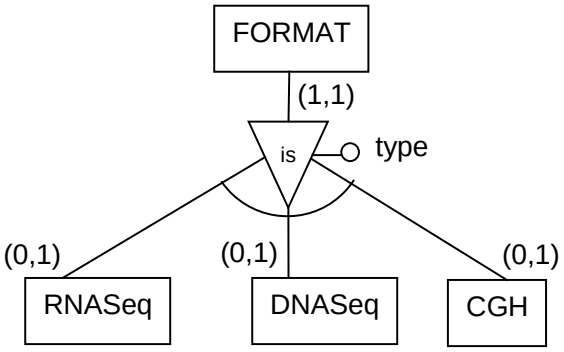

INFOGM(idgm, type, name, description, remarks, histgrade, fase, *parentid*, celltype, biomarks)
CELLLINE(idline, name, origin, organism, description, remarks, receptiondate, morfology, *idgm***)**
CLINICALDATA (**NIP**, document, typedocument, name, lastname1, lastname2, birthdate, sex, age, numhistory, nacionality, remarks, survival, symptom, diagnosis, treatment, diseases, typebiopsia, *idline*)
EXPDETAIL(idexperiment, **idpassage**, **idreplicated**, name, type, condition, description, remarks, receptiondate, typereplicated, *idline*)
EXPDETAILOWNER(*idexperiment*, *idpassage*, *idreplicated*, *idowner*, sdate, edate)
OWNER(idowner, document**,** typedocument**,** nacionality**,** name**,** lastname1**,** lastname2**,** datebirth**,** sex**,** dept**,** position**,** degree**,** tlf, email**,** entity)
FORMAT(idformat, type, name, description, remarks, date, entity, tuser, technology, protocol, mode, method, refvergenome, linkfile, formatfile, numreads, lengthseq, *idexperiment, idpassage, idreplicated*)
FORMATPUBLICATION(*idformat*, *idpublication*)
PUBLICATION(idpublication, idpubmed, idgeo, date, entity, type, doi, link, title, description, remarks)
FORMATRESULT(*idformat*, *idresult*, date)



RESULT(idresult, type, name, fdr, description, remarks, date, linkfile, formatfile)

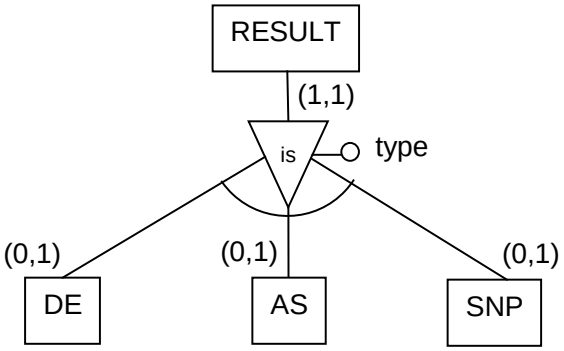
RESULTDE(idresultde, testid, geneid, gene, locus, sample1, sample2, status, value1, value2, lfc, teststat, pvalue, qvalue, significant, result, media, median, stdeviation, rank, *idresult*)

Descripción de jerarquías y generalizaciones del modelo



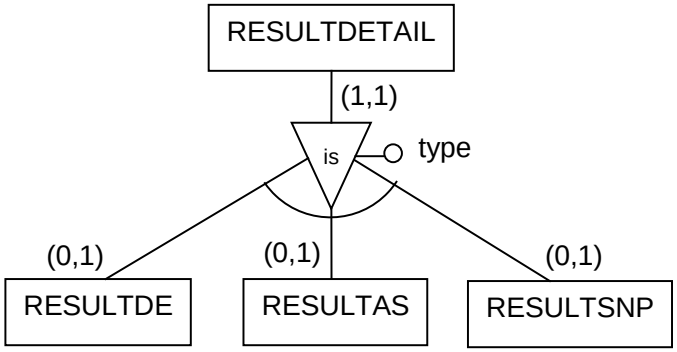
Se trata de una generalización parcial y exclusiva. Representa la clasificación de los distintos tipos de formatos de las muestras que existen en el sistema. Pueden existir otros tipos de formatos no especificados en la jerarquía. Cada formato existe de manera exclusiva en una de las entidades, no puede existir en varias entidades a la vez. El atributo discriminante de la jerarquía es type y tomará uno de los valores de tipos de resultados que existan en el sistema: RNASeq, DNASeq, CGH….

Dado que no hay atributos específicos en los subtipos, se opta por representar en el modelo relacional sólo la entidad FORMAT.



Se trata de una jerarquía parcial y exclusiva. Representa la clasificación de los distintos tipos de ficheros de resultados de análisis realizados sobre las muestras que existen en el sistema. Pueden existir otros tipos de resultados no especificados en la jerarquía. Cada resultado existe de manera exclusiva en una de las entidades, no puede existir en varias entidades a la vez. El atributo discriminante de la jerarquía es type y tomará uno de los valores de tipos de resultados de análisis que existan en el sistema: DE, AS, SNP….

Dado que no hay atributos específicos en los subtipos, se opta por representar en el modelo relacional sólo la entidad RESULT.



Se trata de una jerarquía parcial y exclusiva. Representa la clasificación de los distintos detalles de resultados de los tipos de análisis realizados sobre las muestras que existen en el sistema. Pueden existir otros tipos de detalles de resultados de análisis no especificados en la jerarquía. Cada detalle de resultado existe de manera exclusiva en una de las entidades, no puede existir en varias entidades a la vez. El atributo discriminante de la jerarquía es type y tomará uno de los valores de tipos de resultados de análisis que existan en el sistema: RESULTDE, RESULTAS, RESULTSNP….

Dado el escaso número de atributos comunes y el motivo de mantener una clasificación más detallada de resultados en el sistema, se opta por representar en el modelo relacional sólo las entidades subtipos relevantes, en este caso la entidad RESULTDE, que es la entidad de la que se disponen información.

Descripción de entidades del modelo

A continuación se describen las entidades representadas en el Modelo E/R.

La notación utilizada en la columna Tipo para cada identificador de entidad es la siguiente:

ABREVIATURA	SIGNIFICADO
SP	Simple
OB/OP	Obligatorio/Opcional
UNI	Univaluado
PK	Clave primaria
FK	Clave ajena

◺ **INFOGM**: Almacena información de tipos de Glioma en el sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idgm	PK, SP, OB, UNI	Auto Integer	Identificador de Glioma
type	SP, OB, UNI	Text (68)	Tipo de Glioma (glioblastoma…)
name	SP, OB, UNI	Text (68)	Nombre de Glioma
description	SP, OB, UNI	Text (254)	Descripción de Glioma
remarks	SP, OP, UNI	Text (254)	Observaciones de Glioma

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
histgrade	SP, OP, UNI	Integer	Grado histológico de glioma
fase	SP, OP, UNI	Text (68)	Fase de glioma
celltype	SP, OP, UNI	Text (128)	Tipo de célula de glioma
parentid	FK, SP, OB, UNI	Integer	idgm padre al que pertenece el subtipo
biomarks	SP, OP, UNI	Text (254)	Cadena de biomarcadores conocidos separados por el carácter ;

⩤ **CELLLINE**: Almacena información de las diferentes líneas celulares de trabajo disponibles.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idline	PK, SP, OB, UNI	Auto Integer	Identificador de línea celular
name	SP, OB, UNI	Text (254)	Nombre de línea celular
origin	SP, OB, UNI	Text (32)	Tipo de línea celular (GSC/ Stablished)
organism	SP, OB, UNI	Text (32)	Organismo de línea celular (Human/Mouse)
description	SP, OB, UNI	Text (254)	Descripción de línea celular
remarks	SP, OP, UNI	Text (254)	Observaciones de línea celular
receptionDate	SP, OB, UNI	DateTime	Fecha de recepción de línea celular
morfology	SP, OP, UNI	Text (254)	Descripción de la morfología de línea celular
idgm	FK, SP, OB, UNI	Integer	Identificador de Glioma asociado

⩤ **EXPDETAIL** Almacena información de detalles de experimentos de las distintas líneas celulares del sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idexperiment	PK, SP, OB, UNI	Integer	Identificador de experimento
idpassage	PK, SP, OB, UNI	Integer	Identificador de pase
idreplicated	PK, SP, OB, UNI	Integer	Identificador de replicado
name	SP, OB, UNI	Text (254)	Nombre de detalle experimento
type	SP, OB, UNI	Text (1)	C/T, Control/Treatment
condition	SP, OB, UNI	Text (1)	H/N, Hipoxia/Normoxia
description	SP, OB, UNI	Text (254)	Descripción de detalle experimento
remarks	SP, OP, UNI	Text (254)	Observaciones de detalle experimento
receptiondate	SP, OB, UNI	DateTime	Fecha de recepción de detalle experimento
typereplicated	SP, OP, UNI	Text (1)	Tipo de replicado (T/B, Técnico/Biológico)
idline	FK, SP, OB, UNI	Integer	Identificador de línea celular asociada

⩤ **FORMAT**: Almacena información de formatos de un detalle experimento

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idformat	PK, SP, OB, UNI	Auto Integer	Identificador de formato
type	SP, OB, UNI	Text (128)	Tipo de formato (RNASeq/DNASeq/CGH...)
name	SP, OB, UNI	Text (254)	Nombre de formato
description	SP, OB, UNI	Text (254)	Descripción de formato
remarks	SP, OP, UNI	Text (254)	Observaciones de formato
date	SP, OB, UNI	DateTime	Fecha de creación de formato
entity	SP, OB, UNI	Text (128)	Entidad de creación de formato
tuser	SP, OP, UNI	Text (128)	Nombre de usuario técnico de creación de formato
technology	SP, OP, UNI	Text (128)	Tecnología empleada para la creación de formato (Ejp Sanger/Illumina…)
protocol	SP, OP, UNI	Text (64)	Protocolo utilizado en la creación de formato (Truseq/Deseq…)
mode	SP, OP, UNI	Text (1)	(S/U, Stranded/Unstranded)
method	SP, OP, UNI	Text (2)	(SE/PE, Single End/Paired End)
refvergenome	SP, OP, UNI	Text(16)	Versión del genoma de referencia usada en la creación del formato
linkfile	SP, OP, UNI	Text (254)	Enlace o ubicación de fichero de formato
formatfile	SP, OP, UNI	Text (16)	Formato de fichero de creación de formato
numreads	SP, OP, UNI	Integer	Número de reads que contiene el fichero de formato
lengthseq	SP, OP, UNI	Integer	Longitud de secuencia utilizada para la creación de formato
idexperiment	FK, SP, OB, UNI	Integer	Identificador de experimento
idpassage	FK, SP, OB, UNI	Integer	Identificador de pase
idreplicated	FK, SP, OB, UNI	Integer	Identificador de replicado

◺ **RESULT**: Almacena información de los ficheros de resultados de análisis de formatos del sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idresult	PK, SP, OB, UNI	Auto Integer	Identificador de fichero de resultados
type	SP, OB, UNI	Text (128)	Tipo de fichero de resultados (DE, AS, SNP...)
name	SP, OB, UNI	Text (254)	Nombre de fichero de resultados
fdr	SP, OP, UNI	Integer	Valor de corte utilizado en el análisis relativo a ese fichero
description	SP, OB, UNI	Text (254)	Descripción de fichero de resultados
remarks	SP, OP, UNI	Text (254)	Observaciones de fichero de resultados
date	SP, OB, UNI	DateTime	Fecha de creación de fichero de resultados
linkfile	SP, OP, UNI	Text (254)	Enlace o ubicación del fichero de formato
formatfile	SP, OP, UNI	Text (16)	Formato de fichero de creación de formato

◺ **RESULTDE**: Almacena información de detalles de resultados de análisis de expresión diferencial de formatos del sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idresultde	PK, SP, OB, UNI	Auto Integer	Identificador de resultado de expresión diferencial
testid	SP, OP, UNI	Integer	Identificador único de gen
geneid	SP, OP, UNI	Integer	Nombre genérico o identificador de gen
gene	SP, OP, UNI	Text (64)	Nombre de gen
locus	SP, OP, UNI	Integer	Coordenada genómica
sample1	SP, OB, UNI	Text (254)	Identificador de muestra x
sample2	SP, OB, UNI	Text (254)	Identificador de muestra y
status	SP, OB, UNI	Text (16)	Toma uno de los siguientes valores: OK (prueba exitosa), NOTEST (no hay suficientes alineamientos para la prueba), LOWDATA (demasiado complejo o con poca secuencia), HIDATA (demasiados fragmentos en el locus), o FAIL, cuando una matriz de covarianza errónea u otra excepción numérica impide la prueba.
value1	SP, OB, UNI	Integer	FPKM del gen en la muestra x
value2	SP, OB, UNI	Integer	FPKM del gen en la muestra y
lfc	SP, OP, UNI	Integer	Logaritmo en base 2 del valor Fold Change (y/x)
teststat	SP, OB, UNI	Integer	El valor del estadístico t utilizado para calcular la importancia del cambio observado en FPKM
pvalue	SP, OB, UNI	Integer	pvalor no corregido del estadístico t
qvalue	SP, OP, UNI	Integer	pvalor FDR ajustado del estadístico t
significant	SP, OP, UNI	Text (3)	Puede tomar el valor YES/NO, dependiendo de si pvalue es mayor que el FDR después de la corrección por multiple-testing Benjamini-Hochberg
result	SP, OP, UNI	Text (16)	Puede tomar el valor Up/Down según el resultado de la expresión diferencial sea upregulado o downregulado en la muestra 2 en función del FDR definido en el fichero de resultados
rank	SP, OP, UNI	Integer	Número de orden según valor LFC ascendente
idresult	FK, SP, OB, UNI	Integer	Identificador de fichero de resultados

◺ **CLINICAL DATA**: Almacena información de los datos clínicos de las líneas celulares del sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
NIP	PK, SP, OB, UNI	Auto Integer	Identificador de paciente
document	SP, OB, UNI	Integer	Número de identificación personal
typedocument	SP, OB, UNI	Text (16)	NIF/NIE
name	SP, OB, UNI	Text (128)	Nombre de paciente
lastname1	SP, OB, UNI	Text (128)	Primer apellido de paciente
lastname2	SP, OP, UNI	Text (128)	Segundo apellido de paciente
birthdate	SP, OP, UNI	DateTime	Fecha de nacimiento de paciente
sex	SP, OB, UNI	Text (1)	Genero de paciente (M/F)
age	SP, OP, UNI	Integer	Edad de paciente
numhistory	SP, OB, UNI	Integer	Número de historia clínica de paciente
nacionality	SP, OP, UNI	Text (128)	Nacionalidad de propietario
remarks	SP, OP, UNI	Text (254)	Observaciones clínicas
survival	SP, OP, UNI	Text (1)	Supervivencia (Y/N) de paciente
sympton	SP, OP, UNI	Text (254)	Descripción de síntoma de paciente
diagnosis	SP, OP, UNI	Text (254)	Diagnóstico de paciente
treatment	SP, OP, UNI	Text (254)	Tratamiento
diseases	SP, OP, UNI	Text (254)	Otras enfermedades de paciente
typebiopsia	SP, OP, UNI	Text (1)	Tipo de biopsia S/L, Sólida/Líquida
idline	FK SP, OB, UNI	Integer	Identificador de línea asociada

◣ **PUBLICATION**: Almacena información de publicaciones realizadas de formatos.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idpublication	PK, SP, OB, UNI	Auto Integer	Identificador de publicación
idpubmed	SP, OB, UNI	Integer	Identificador de publicación en pubmed
idgeo	SP, OP, UNI	Integer	Identificador de publicación en geo
date	SP, OB, UNI	DateTime	Fecha de publicación
entity	SP, OB, UNI	Text (128)	Entidad en la que se ha publicado
type	SP, OP, UNI	Text (128)	Tipo de publicación
doi	SP, OP, UNI	Text (254)	Identificador digital de la publicación
link	SP, OB, UNI	Text (254)	Enlace a la publicación
title	SP, OB, UNI	Text (254)	Título del artículo publicado
description	SP, OB, UNI	Text (254)	Descripción de la publicación
remarks	SP, OP, UNI	Text (254)	Observaciones

◣ **OWNER**: Almacena información de usuarios propietarios de detalles de experimentos

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idowner	PK, SP, OB, UNI	Auto Integer	Identificador de propietario
document	SP, OB, UNI	Integer	Número de identificación personal
typedocument	SP, OB, UNI	Text (16)	NIF/NIE
nacionality	SP, OP, UNI	Text (128)	Nacionalidad de propietario
name	SP, OB, UNI	Text (128)	Nombre de propietario
lastname1	SP, OB, UNI	Text (128)	Primer apellido de propietario

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
lastname2	SP, OP, UNI	Text (128)	Segundo apellido de propietario
datebirth	SP, OP, UNI	DateTime	Fecha de nacimiento de propietario
sex	SP, OB, UNI	Text (1)	Genero de paciente (M/F)
dept	SP, OP, UNI	Text (128)	Departamento de propietario
position	SP, OP, UNI	Text (128)	Puesto de trabajo de propietario
degree	SP, OP, UNI	Text (128)	Graduación de propietario
tlf	SP, OB, UNI	Text (64)	Teléfono de contacto de propietario
email	SP, OB, UNI	Text (128)	Dirección email de propietario
entity	SP, OP, UNI	Text (128)	Entidad del propietario

1.6 Descripción de entidades provenientes de interrelaciones

FORMATRESULT: Almacena información de resultados asociados a formatos.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idformat	PK, FK, SP, OB, UNI	Integer	Identificador de formato
idresult	PK, FK, SP, OB, UNI	Integer	Identificador de fichero de resultados
date	PK, FK, SP, OB, UNI	DateTime	Fecha de obtención de fichero de resultados

FORMATPUBLICATION: Almacena información de las publicaciones asociados a formatos.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idformat	PK, FK, SP, OB, UNI	Integer	Identificador de formato
idpublication	PK, FK, SP, OB, UNI	Integer	Identificador de publicación

EXPDETAILOWNER: Almacena información de propietarios asociados a detalles de experimentos del sistema.

ATRIBUTO	TIPO	FORMATO	DESCRIPCION
idexperiment	PK, FK, SP, OB, UNI	Integer	Identificador de experimento
idpassage	PK, FK, SP, OB, UNI	Integer	Identificador de pase
idreplicated	PK, FK, SP, OB, UNI	Integer	Identificador de replicado
idowner	PK, FK, SP, OB, UNI	Integer	Identificador de propietario
sdate	PK, SP, OB, UNI	DateTime	Fecha de inicio de propiedad
edate	PK, SP, OB, UNI	DateTime	Fecha fin de propiedad

ANEXO 2. FORMATO DE FICHERO DIFF

Definición de columnas del fichero de expresión diferencial en formato diff

NÚMERO COLUMNA	NOMBRE COLUMNA	EJEMPLO	DESCRIPCIÓN
1	tested id	A1BG	Identificador único de gen
2	gene id	A1BG	Nombre genérico o identificador de gen
3	gene	A1BG	Nombre de gen
4	locus	chr19:58858171-58874214	Coordenada genómica
5	sample1	CS123NGR	Identificador de muestra x
6	sample2	CS101112NGR	Identificador de muestra y
7	status	OK	Toma uno de los siguientes valores: OK (prueba exitosa), NOTEST (no hay suficientes alineamientos para la prueba), LOWDATA (demasiado complejo o con poca secuencia), HIDATA (demasiados fragmentos en el locus), o FAIL, cuando una matriz de covarianza errónea u otra excepción numérica impide la prueba.
8	value1	0.0222979	FPKM del gen en la muestra x
9	value2	3.15227	FPKM del gen en la muestra y
10	log2(fold change)	7.14334	Logaritmo en base 2 del valor Fold Change (y/x)
11	teststat	2.02369	El valor del estadístico t utilizado para calcular la importancia del cambio observado en FPKM
12	pvalue	0.2506	pvalor no corregido del estadístico t
13	qvalue	0.379358	pvalor FDR ajustado del estadístico t
14	significant	no	Puede tomar el valor yes/no, dependiendo de si pvalue es mayor que el FDR después de la corrección por multiple-testing Benjamini-Hochberg